REMARKS

Claims 1-25, 27-32, and 44 are pending; claims 1, 9, 14, 19, and 27 have been amended.

Claim 1, 9, 14, and 19 have been amended to more particularly point out and distinctly claim Applicant's invention. Support for this amendment can be found throughout the specification, for example, at page 31, line 11 through page 34, line 12.

Claim 27 has been amended to conform the claim with claim 24, from which it depends. Support for this claim amendment can be found throughout the specification, for example, at page 9, lines 8-9.

The claims have also been amended to correct minor typographical errors.

The rejections of the claims under 35 U.S.C. §§102 and 103 have been withdrawn.

The above amendments of the claims are done without prejudice to further prosecution of other embodiments of this invention in a continuation, continuation-in-part, divisional, or other related application. None of the above amendments adds any new matter to the Application.

I. Priority Application

The Office Action has denied the Application the benefit of the May 14, 1998 filing date of the priority document because "as presently written, the full scope embraced by each claim was not disclosed in the provisional application." (Office Action, page 2). Specifically, the Office Action states that "the provisional application is essentially a research paper and discloses particular experiments performed. There is no generic disclosure of the methods as presently claimed." (Office Action, page 2).

Applicant respectfully submits that she is entitled to priority to the May 14, 1998, date for as much as what is disclosed and fully supported by the priority document.

The priority document teaches that normal huntingtin was associated with MLK2, that expression of MLK2 activated the SEK1-JNK pathway and induced apoptosis in neuronal cells, and that co-expression of MLK2 with mutated huntingtin induced toxicity in non-neuronal cells (293 embryonic kidney cells) while co-expression of MLK2 with normal huntingtin did not (see priority document at page 2). The priority document teaches that neuronal toxicity was induced

by either mutated huntingtin or by MLK2, and that this toxicity could be attenuated by a dominant negative SEK1 (see priority document at page 8). The priority document later states, "normal huntingtin is associated with MLK2 in intact cells and such association dos not generate any cell toxicity" (priority document at page 9).

Applicant avers that upon reading the specification, the ordinarily skilled artisan would understand that (1) activated MLK2 activity induced neuronal cell toxicity; (2) MLK2 activity and its ability to induce cell death can inhibited by co-expression of normal huntingtin. Applicant avers that upon reading the specification, the ordinarily skilled artisan would understand that a compound (such as normal huntingtin) has an ability to inhibit neuronal cell death if a neuronal cell with activated MLK activity does not die when contacted with the compound since MLK is a constitutively active kinase whose activity is held in check only because MLK associates with normal huntingtin (see priority document at pages 6-7 and Figures 4B and 4C). The priority document teaches the ordinarily skilled artisan that compounds other than normal huntingtin can be similarly assessed for their ability to inhibit neuronal cell death. Thus, the priority document discloses the full scope of the cell culture based methods of the claimed invention.

Furthermore, Applicant posits that the priority document also discloses the full scope of the cell free based methods of the claimed invention. As described in the priority document at page 5, MLK is a protein kinase that directly binds to and activates SEK1. Expression of mutant huntingtin in MLK-expressing neuronal cells induced cell death; however, when a dominant negative form of SEK1 was co-expressed with mutated huntingtin in MLK-expressing neuronal cells, death of the cells was blocked (see priority document at page 5). Thus, one of ordinary skill in the art, upon reading the priority document, would understand that active MLK's kinase activity in binding to and activating SEK1 results in cell death—accordingly, the ordinarily skilled artisan would understand that a compound (such a normal huntingtin) that inhibits MLK kinase activity and therefore inhibits phosphorylation and activation of a MLK substrate (*e.g.*, SEK1) would also inhibit neuronal cell death. Accordingly, Applicant avers that the priority document discloses the full scope of the cell free based methods of the claimed invention.

Given the broad teachings of the priority document, and the knowledge of the ordinarily skilled artisan as of the priority document's filing date, Applicant avers that the Application is entitled to the May 14, 1998 filing date of the priority document, since the priority document disclosed the full scope of the claimed invention.

II. Rejection Under 35 U.S.C. §112, first paragraph.

Claims 1-25, 27-32, and 44 stand rejected under 35 U.S.C. §112, first paragraph, because the specification "does not reasonably provide enablement for the breadth of the claims as written" (Office Action, page 3). Specifically, the Office Action states that "the specification does not describe nor enable performing the claimed methods in an intact mammal." (Office Action, page 4).

Applicant has overcome this ground for rejection by amendment to the claims, as supported by the remarks made below.

As an initial matter, Applicant has amended claims 1, 9, 14, and 19 to clarify that the cell culture based methods are what was intended to be covered by these claims. This being the case, Applicant respectfully avers that the specification fully enables the claims as presently amended because the ordinarily skilled artisan would have known, upon reading the specification, that the claimed methods allow an assessment of a compound's ability to inhibit MLK activity, and therefore inhibit neuronal cell death. As to claim 24, and the claims dependent thereon, which cover *in vitro* cell-free methods, Applicant likewise avers that the specification fully enables the ordinarily skilled artisan to practice these claimed cell-free methods to assess a compound's ability to inhibit MLK activity, and therefore inhibit neuronal cell death.

As to the assertion that the claimed methods allow an assessment of compounds that inhibit (and not necessarily prevent) neuronal cell death, where death (or prevention thereof) of a neuronal cell is concerned, Applicants avers that the ordinarily skilled artisan would understand that "inhibit" and "prevent" are words that can be used interchangeably—any compound assessed as having an ability to inhibit cell death would also prevent cell death, and vice versa. The specification has described, in Example I (specification, page 24, line 21 through page 27, line 13), a cell culture based method for studying an exemplary neurodegenerative disease,

namely Huntington's disease. Applicant avers that the ordinarily skilled artisan, upon reading Example I, would understand that any compound assessed by the methods of the invention as having an ability to inhibit death of cultured neuronal cells induced to apoptose by expression of polyglutamine-expanded huntingtin would also inhibit death of *in vivo* neuronal cells induced to apoptose by expression of polyglutamine-expanded huntingtin.

Moreover, Applicant avers that the ordinarily skilled artisan would understand that using the methods of the invention, an assessment may be made of a compound's ability to inhibit neuronal cell death in general, regardless of whether or not the neuronal cell death is associated with Huntington's disease, Alzheimer's disease, or any other neurological condition. That the specification lists many different types of neurological conditions (see specification at page 13, line 6 through page 14, line 2) merely evidences that the claimed invention is meant to embrace any and all of these diseases, with Huntington's disease and Alzheimer's disease being but examples of the types of neurological conditions covered by the claims.

The Office Action further states that because the specification "does not appear to specifically define the metes and bound of the intended activities [sic], proteins, and activities", "the specification does not describe or enable identification of any other MLK proteins or activities meeting the functional limitations of the claims and it is deemed to constitute undue experimentation to determine them." (Office Action, page 5).

Applicant avers that the specification has provided sufficient guidance for identifying MLK proteins and activities for use in the claimed invention.

First of all, Applicant notes that the Application does not claim MLK proteins, MLK nucleic acid molecules, or methods of producing MLK proteins or nucleic acid molecules. Rather, as is explained below, the Application claims using an MLK protein in a method for preventing neuronal cell death, where one of ordinarily skill in the art, without undue experimentation, could identify an MLK protein based upon the teachings of the specification. Accordingly, the case cited by the Office Action, In re Maizel, 27 USPQ2d 1662 is not appropriately applied since In re Maizel concerns the lack of enablement of a claim covering DNA and vectors containing the DNA.

As to identifying MLK proteins useful in the claimed invention, Applicant respectfully avers that such identification would be routine to one of ordinary skill in the art upon reading the specification. To buttress her position, Applicant respectfully directs the Examiner's attention to page 499, left column, of Dorow et al., *Eur. J. Biochem.* 234: 492-500 (1995) ("Dorow"; provided herewith as Appendix A; incorporated by reference into the specification (see page 3, line 2 and page 24, lines 16-17)). There, Dorow teaches that MLK2 expression is observed in brain, skeletal muscle, and in the pancreas, and that MLK3 expression is observed in most cell lines and tissues examined. Moreover, Applicant notes that MLK2, when co-expressed with mutated huntingtin, induced apoptosis in embryonic kidney cells (see specification at page 33, lines 15-19).

Thus, while the specification states at page 10, line 13 that MLK2 "is a neuronal form of MLKs" (emphasis added), Applicant avers that the ordinarily skilled artisan would understand that MLK2 may not be the only MLK in neuronal cells, particularly given the teaching by Dorow that MLK2 is also found in colonic cells. The specification teaches that MLKs "are the only known kinases that directly activate the SEK1-JNK cascade and contain a SH3 domain as well as a SH3 domain binding site." (page 10, lines 11-12) Given this teaching, Applicant avers that the skilled artisan would understand that a kinase which (1) contains an SH3 domain, (2) contains an SH3 domain binding site, and (3) directly activates the SEK1-JNK cascade is a MLK within the scope of the claims. Given the teachings of the specification that demonstrate that an examplary MLK, MLK2, directly binds to and phosphorylates SEK1 (see, *e.g.*, page 36, lines 4-28), no further undue experimentation is required on the part of the ordinarily skilled artisan to make the determination of which MLK proteins are encompassed by the scope of the claims.

Additionally, Applicant avers that one of ordinary skill in the art would understand what the metes and bounds of intended activities of MLK are. The specification, at page 9, lines 8-10, states that MLK is activated by being bound by an SH3 domain on a triggering protein. Activated MLK then directly binds to and stimulates a SEK1 protein. While it is true that activated MLK has enzymatic activity (*e.g.*, kinase activity), activated MLK also has other activities, including activated MLK's ability to bind SEK1 (see, *e.g.*, page 12, lines 24-25). Those of ordinary skill at the time the Application was filed will understand that an "activated"

protein need not be enzymatically active, and a protein that is enzymatically active may also be activated by a conformation change revealing or hiding a regulatory domain. Thus, MLK protein activity includes, without limitation, a kinase activity and an ability to bind SEK1. Applicant avers that the ordinarily skilled artisan would understand that it matters not how MLK is activated to induce apoptosis of neuronal cell, only that when MLK is activated (in the presence of appropriate stimuli), neuronal cell death results. Of course, activated MLK activity will increase or decrease depending upon increased or decreased rates of MLK transcription and/or translation (see specification, page 12, lines 20-21).

Given its teachings, Applicant avers that the specification does not merely provide an invitation to experiment analogous to that in Genentech Inc. v. Novo Nordisk A/S, 42 USPQ2d 1001, as has been asserted by the Office Action (see page 6). Rather, Applicant avers that the specification describes a model for neuronal cell death and teaches methods for utilizing an MLK protein to assess the ability of compounds to inhibit neuronal cell death. As such, Applicant avers that the specification has fulfilled the requirements of 35 U.S.C. §112, first paragraph.

The Office Action has stated that a "reasonable correlation must exist between the scope of the claims and the scope of enablement set forth." (Office Action, page 5). Applicant does not dispute this requirement but, rather, posits that the specification has met the requirements for enablement. To buttress her position, Applicant respectfully the Examiner's attention the specification at page 8, lines 23-28. There, the specification teaches that inhibition of MLK2 can protect a neuronal cell from apoptosis induced by polyglutamine-expanded huntingtin. As taught by the reference, Huntington's Disease Collaborative Research Group, *Cell* 72:971-983 (1993) (provided herewith as Appendix B), which is cited by the specification at page 8, line 28 and incorporated by reference (see specification page 24, lines 16-17), expression of polyglutamine-expanded huntingtin caused Huntington's Disease in humans. As the specification summarizes at page 32, line 15, "MLK-associated activity was involved in neuronal loss in Huntington's diseases."

The specification teaches, at page 24, line 21 through page 27, line 13, a cell culture based model for studying polygluatmine-expanded huntingtin induced neurodegeneration was developed. Later, at page 31, line 11 through page 34, line 12, the specification teaches that an

examplary MLK protein, MLK2, is associated with huntingtin and that a mutant MLK2 protein lacking kinase activity blocks apoptosis induced by glutamate of kainate receptor activation in neuronal cells. Still later, at page 34, line 14 through page 35, line 4 and page 36,line 3 through page 37, line 22, the specification describes the development of a cell free based model for studying neurodegeneration.

Applicant avers that the specification as filed has met the standards of enablement as set forth in Manual of Patent Examining Procedures §2164.02 (7th Edition, Rev. 1, Feb. 2000). There, under the section "Correlation: in vitro/in vivo", MPEP §2164.02 states "if the art is such that a particular model is recognized as correlating to a specific condition, then it should be accepted as correlating unless the examiner has evidence that the model does not correlate. Even with such evidence, the examiner must weight the evidence for an against correlation and decide whether one skilled in the art would accept the model as reasonably correlating to the condition." Applicant posits that one of ordinary skill in the art would have understood that the *in vitro* cell culture model and cell free model for neurodegenerative disease set forth in the specification correlates with *in vivo* neurological conditions, including those described in the specification (see, *e.g.*, page 13, line 6 through page 14, line 2). The standards for enablement of 35 U.S.C. §112, first paragraph, require nothing more.

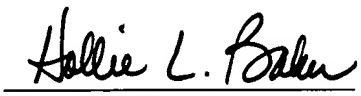
Accordingly, Applicant respectfully requests the grounds for this 35 U.S.C. §112, first paragraph, be reconsidered and withdrawn.

III. Conclusion.

Applicants posit that the presently maintained rejections of the pending claims have been fully overcome by amendment and/or argument. Accordingly, Applicants respectfully submit that the pending claims are in condition for allowance. If the Examiner believes that any further discussion of this communication would be helpful, she is encouraged to contact the undersigned by telephone.

Aside from the fee for the two month extension of time, no fees are believed to be due in connection with this communication. However, please apply any additional charges, or credit any overpayment, to our Deposit Account No. 08-0219.

Respectfully submitted,



Hollie L. Baker
Registration No. 31,321

Hale and Dorr LLP
60 State Street
Boston, MA 02109
617 526-6110 (Telephone)
617 526-5000 (Fax)

Dated: November 30, 2000

Appendix A

Eur. J. Biochem. 234, 492–500 (1995)

Complete nucleotide sequence, expression, and chromosomal localisation of human mixed-lineage kinase 2

Donna S. DOROW[1], Lisa DEVEREUX[1], Guo-fen TU[2,3], Gareth PRICE[1], Jillian K. NICHOLL[4], Grant R. SUTHERLAND[4] and Richard J. SIMPSON[2,3]

[1] Research Division, The Peter MacCallum Cancer Institute Melbourne, Victoria, Australia
[2] The Joint Protein Structure Laboratory, Ludwig Institute for Cancer Research (Melbourne Branch), Victoria, Australia
[3] The Walter and Eliza Hall Institute for Medical Research Parkville, Victoria, Australia
[4] Center for Medical Genetics, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, Adelaide, Australia

(Received 12 June 1995) – EJB 95 0959/3

Protein kinases play pivotal roles in the control of many cellular processes. In a search for protein kinases expressed in human epithelial tumour cells, we discovered two members of a novel protein kinase family [Dorow, D. S., Devereux, L., Dietzsch, E. & de Kretser, T. A. (1993) *Eur. J. Biochem. 213*, 701–710]. Due to the unique mixture of structural domains within their amino acid sequences, we named the family mixed-lineage kinases (MLK). We initially isolated clones encoding partial cDNAs of MLK1 and 2 from a human colonic cDNA library. The MLK2 cDNA was subsequently used to screen a human brain cDNA library and we have now cloned and sequenced a 3454-bp cDNA encoding the full-length MLK2 protein. The predicted MLK2 polypeptide has 954 amino acids and contains a *src* homology 3 (SH3) domain, a kinase catalytic domain, a double leucine zipper and basic domain, and a large C-terminal domain. The 22-amino-acid N-terminal region has four glutamic acid residues immediately following the initiator methionine. Beginning at amino acid 23, the 55-amino-acid SH3 domain contains a 5-amino-acid insert in a position corresponding to inserts of 6 and 15 residues in the SH3 domains of *n-src* and the phosphatidylinositol 3′-kinase. Adjacent to the SH3 domain is a kinase catalytic domain with conserved motifs associated with both serine/threonine and tyrosine specificity. Beginning nine residues C-terminal to the catalytic domain, there are two leucine/isoleucine zippers separated by a 13-amino-acid spacer sequence and followed by a stretch of basic residues. The polybasic sequence contains a motif that is similar to nuclear localisation signals from several proteins. The C-terminal domain is composed of 491 amino acids of which 17% are serine or threonine and 16% are proline. This domain also has a biased ratio of basic to acidic amino acids with a calculated pI of 9.38. When used as a probe to examine mRNA expression in human tissues, a MLK2 cDNA hybridised to a species of 3.8 kb that was expressed at highest levels in RNA from brain and skeletal muscle. The 3454-bp cDNA was also used for fluorescence *in situ* hybridisation to localise the MLK2 gene to human chromosome 19 q13.2.

Keywords: protein kinase; mixed-lineage kinase; leucine zipper; SH3 domain; DNA sequence.

Protein kinases play critical roles in the regulation of biochemical and morphological changes associated with cellular growth and division (D'Urso et al., 1990; Birchmeier et al., 1993). They serve as growth factor receptors and signal transducers and have been implicated in cellular transformation and malignancy (Hunter and Karin, 1992; Posada and Cooper, 1992; Hunter and Pines, 1994). Protein kinases can be divided into two main groups both by amino acid sequence similarity and specificity for either serine/threonine or tyrosine. A small number of dual-specificity kinases are structurally like the serine/threonine-specific group. Within the broad classifications, kinases can be further sub-divided into families whose members share a higher degree of catalytic domain amino acid sequence identity and also have similar biochemical properties (Hanks et al., 1988). Most protein kinase family members also share structural features outside the kinase domain that reflect their particular cellular roles. These include regulatory domains that control kinase activity or interaction with other proteins (Hanks, 1991). Two regulatory elements, originally identified as conserved sequences in members of the *src*-related kinase family, are the *src* homology 2 (SH2) and 3 (SH3) domains (Sadowski et al., 1986; Koch et al., 1991). These domains have now been found in a variety of proteins involved in intracellular signalling pathways where they link activated cell surface receptors to downstream effectors (Pawson and Gish, 1992). SH3 domains are also found in cytoskeletal proteins (Drubin et al., 1990; Rodaway et al., 1989). While SH2 domains bind phosphorylated tyrosines in the cytoplasmic domains of activated receptors, SH3 domains bind proline-rich sequences in their target molecules. Additional roles have been suggested for SH3 domains including localisation of proteins to the vicinity of the cell membrane where the early events in signalling occur (Booker et al., 1993; Rodaway et al., 1989; Bar-Sagi et al., 1993). Furthermore, it has recently been shown that SH3 domains also participate in regulating the activ-

Correspondence to D. S. Dorow, Research Division, The Peter MacCallum Cancer Institute Melbourne, Victoria 3000 Australia
Fax: +61 3 9656 1411.

Abbreviations. MLK, mixed-lineage kinase(s); SH, *src* homology domain; p85-PtdIns3K, p85 subunit of the phosphatidylinositol 3′-kinase; FISH, fluorescent *in situ* hybridisation.

Note. The novel nucleotide sequence data published here have been deposited with the EMBL sequence databank and are available under accession number X90846.

ity of both protein kinases (Superti-Furga et al., 1993) and guanosinetriphosphatase effector proteins (Gout et al., 1993).

Another regulatory domain, usually found in transcription factors such as the oncogenes *fos*, *jun*, and *myc*, is the leucine zipper (Landschultz et al., 1988). Leucine zipper sequences, containing a leucine or isoleucine residue at every seventh position for a stretch of at least 22 amino acids, take up amphipathic helical conformations with the leucine side chains forming a stripe down one face. In the transcription factors, the leucine zipper is preceded by a stretch of basic amino acids that constitute the DNA-binding region. Leucine zippers promote dimerization though hydrophobic interactions between heptad leucines (O'Shea et al., 1991). Such dimerization appears to activate DNA-binding by orientating the basic side chains of DNA-binding residues to enable correct contact with DNA (Vinson et al., 1989). While leucine zippers are not commonly associated with protein kinases, the cyclic-GMP-dependent kinase has a leucine zipper through which it forms its active state dimer (Wolfe et al., 1989).

In a previous report (Dorow et al., 1993), we described a novel protein kinase family, the mixed-lineage kinases (MLK). These kinases have an unusual catalytic domain structure that is a hybrid between the tyrosine and serine/threonine-specific types. In addition, they posses a unique double leucine zipper and basic domain that has only been found in members of the MLK family. We first reported partial amino acid sequences for two members of this family, MLK1 and 2. Further studies revealed that each of these proteins contains a SH3 domain (Dorow et al., 1994; Dorow, D., unpublished results). Recently, two further members of the MLK family have been reported. One of these, MLK3 (Ing et al., 1994), also reported as PTK1 (Ezoe et al., 1994) and SPRK (Gallo et al., 1994), is very closely related to MLK1 and 2. A fourth, more distantly related member of the MLK family, DLK (Holzman et al., 1994), contains MLK-like catalytic and double leucine zipper domains but lacks a SH3 domain. Thus, the MLK enzymes are an emerging family of protein kinases with a unique mixture of structural domains. In addition to the unusual nature of their kinase and double leucine zipper domains, they are the only protein kinases thus far reported that contain SH3 domains in the absence of SH2 domains. In the present study, we describe the complete nucleotide sequence, tissue expression, and chromosomal localisation of MLK2 from human brain.

MATERIALS AND METHODS

Cloning and sequence analysis. Segments of cDNAs encoding catalytic subdomains of protein kinases expressed in the epithelial tumour cell line Colo 16 (Moore et al., 1975) were amplified from RNA by reverse transcriptase PCR by the method of Goblet et al. (1989). Degenerate PCR primers were based on sequences encoding conserved motifs in subdomains VIb and VIII (Wilks, 1989) of the epidermal-growth-factor receptor family kinase catalytic domains (Hanks et al., 1988). Sequences of the primers were as follows: forward primer (with an added *Bam*H1 site underlined) 5'-CGGATCCGTG(A)CAC-C(A)GT(CG)G(A)ACC(T)T; reverse primer (with an added *Eco*R1 site underlined) 5'-GGAATTCACCA(G)TAA(G)CT-CCA-G(C)ACATC. Several PCR products were cloned into M13 and sequenced using a T7 Super-Base sequencing kit (Bresatec). One 216-bp PCR product was used as a probe to screen a human colon λgt 11 cDNA library (Clontech). The library was made from normal tissues excised from around the colon cancer of a 53-year-old male (Clontech library catalogue). Four clones were isolated and their inserts sequenced. All of these clones represented overlapping areas of the same sequence that we designated MLK1. One cDNA from this screen was used as a probe to rescreen the same library, and a 1034-bp cDNA was isolated and sequenced (Dorow et al., 1993). The 1034-bp MLK2 cDNA was then used as a probe to screen a human brain λgt 10 library. Approximately 0.5×10^6 clones were screened over several screenings, and one 3454-bp clone was isolated. The insert from this clone was subcloned into pUC18 and sequenced on both strands using an Applied Biosystems 373 automated DNA sequencer. Sequencing reactions were carried out with a Prism Ready Reaction dyedeoxy terminator cycle sequencing kit (Applied Biosystems) according to manufacturer's instructions. Temperature cycling was performed on a Perkin Elmer GeneAmp PCR system 9600. The cycling procedure was 15 s at 93°C, 15 s at 50°C, and 4 min at 60°C for 25 cycles. The procedure was modified for (G+C)-rich templates to include 5% (by vol.) dimethyl sulfoxide and a denaturing temperature of 98°C. All chemicals were purchased from Sigma unless otherwise stated.

Northern-blot analysis. A multi-tissue Northern (MTN) blot (2 μg mRNA/lane) was purchased from Clontech and treated according to standard procedures supplied by the manufacturer. Briefly, the blot was prehybridised in 5×NaCl/P$_i$/EDTA (0.75 M NaCl, 0.05 M NaH_2PO_4 and 5 mM EDTA), 10×Denhardt's solution [1% (mass/vol.) Ficoll (Pharmacia type 400), 1% (mass/vol.) polyvinylpyrrolidone, and 1% (mass/vol.) BSA], 100 μg/ml sheared, denatured salmon sperm DNA, 50% (by vol.) deionised formamide (Fluka) and 2% (mass/vol.) SDS for 4 h at 42°C. Probes were labelled with [^{32}P]dA or CTP (NEN Dupont, 3000 Ci/mM) by random priming (Feinberg and Vogelstein, 1983) to a specific activity of $\approx10^9-10^9$ cpm/μg, added to the filters in prehybridisation solution, and incubated overnight at 42°C. Final stringency washes were in 0.5×NaCl/Cit (7.5 mM $Na_3C_6H_5O_7$, pH 7.0, 75 mM NaCl) and 0.5% (mass/vol.) SDS at 65°C.

Fluorescence *in situ* hybridisation. The 3454-bp MLK2 cDNA was labelled by nick-translation with biotin-14-d[ATP] (Gibco BRL) and hybridised *in situ* at a concentration of 10 ng/μl to metaphase chromosomes from two normal male donors. Chromosomes were prepared from peripheral blood leucocytes by standard procedures following the method of Wheater and Roberts (1987). Fluorescence *in situ* hybridisation (FISH) was performed as described by Callen et al. (1990) except that the chromosomes were stained before analysis with both propidium iodide and 4',6-diamidino-2-phenylindole. Images of metaphase preparations were recorded using a Panasonic WV-BL600 CCD camera.

RESULTS

Cloning and nucleotide sequence analysis of MLK2. In a search for protein kinases expressed in human epithelial tumour cells, we used a PCR strategy to amplify segments of kinase catalytic domains from epithelial cell RNAs. The primers were cDNAs encoding conserved motifs in the catalytic domains (Wilks, 1989) of epidermal-growth-factor receptor family members. Several PCR products were cloned and sequenced. One PCR product that represented a novel kinase catalytic domain sequence was used to probe a human colonic cDNA library and several clones were isolated. These cDNAs encoded overlapping areas of a putative protein kinase that we named MLK1 (Dorow et al., 1993). When the MLK1 cDNA was used to rescreen the same library, a 1034-bp cDNA fragment that had 65% nucleotide sequence identity to MLK1 was isolated. It was clear that together these two molecules represented a new family of protein kinases.

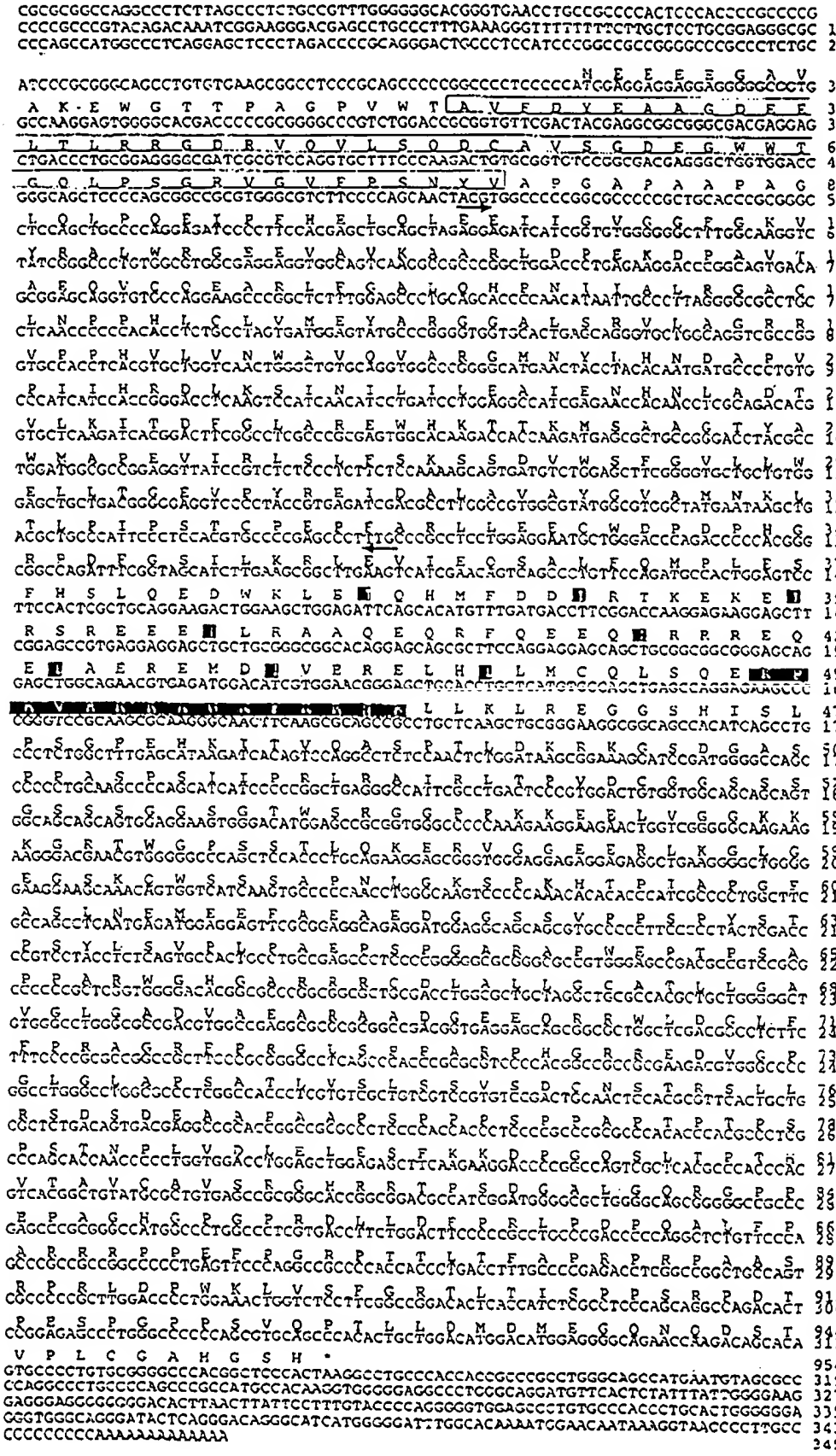


Fig. 1. The nucleotide and deduced amino acid sequences of the human MLK2 cDNA. Nucleotide and amino acid numbers are indicated at the right-hand side of each line. The SH3 domain amino acids are enclosed in a non-shaded box and the beginning and end of the kinase catalytic domain are delineated above the line (— —). Amino acids that define the two leucine zipper heptads and the basic domain are enclosed in black boxes. The in-frame termination codon is also indicated (*).

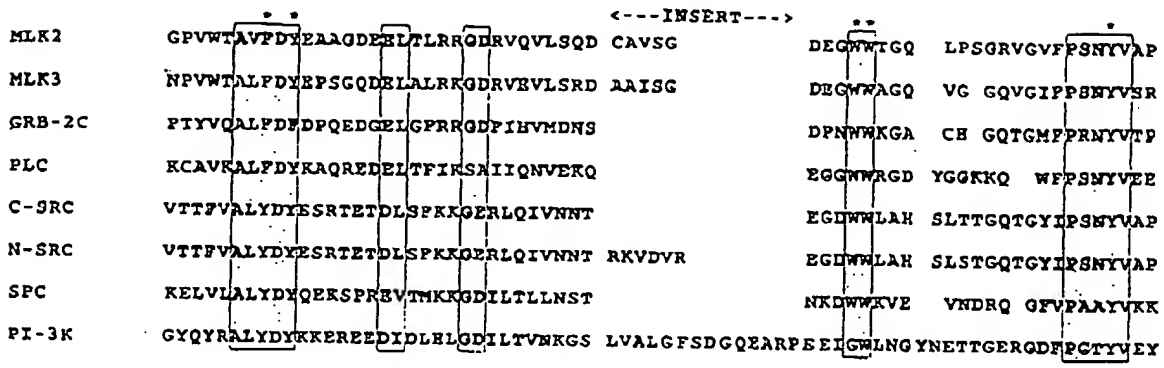


Fig. 2. Alignment of SH3 domain amino acid sequences. The predicted amino acid sequence of the SH3 domain of MLK2, aligned with the sequences of MLK3, the C- terminal SH3 domain of GRB 2 (GRB 2C), phospholipase Cγ (PLC), *c-src*, *n-src*, spectrin (SPC), and p85-PtdIns3K (PI-3K). Amino acids that make up the SH3 domain consensus motifs are shaded and the positions of the conserved aromatic amino acids are indicated (*).

The 1034-bp MLK2 cDNA fragment was then used as a probe to screen a human brain cDNA library and a single clone was isolated and its insert sequenced. The 3454-bp nucleotide sequence of this insert is shown in Fig. 1. The insert contains 289 bp of 5′-untranslated nucleotides, an open reading frame of 2862 bp, and 304 bp of 3′-untranslated nucleotides. The putative initiator AUG codon begins at nucleotide 290 and is preceded by an in-frame stop codon beginning at nucleotide 122. This methionine codon is contained within the sequence CCCAUGG (positions −3 to +4). According to the scanning model of Kozak (1989), translation is usually initiated at the most 5′ AUG in a favourable context. The most favourable sequence for recognition by eukaryotic ribosomes is A/GCCAUGG (positions −3 to +4). Mutagenic analysis of this sequence, however, has shown that in the absence of the purine at the −3 position, there is a strong preference for recognition of sequences with a G at position +4 (Kozak, 1986). The sequence surrounding the AUG beginning at nucleotide 290 fulfils this requirement. There is one upstream AUG, beginning at nucleotide 165, in a more favourable context for initiation. This AUG, however, is in a different frame from that of the main open reading frame of the sequence and is followed by an in-frame stop codon beginning at nucleotide 255. In such a situation, the main start site of the protein is predicted to be the downstream AUG (Kozak, 1986). Thus, the AUG at nucleotides 290−292 most likely encodes the actual N-terminus of the MLK2 protein. At the extreme 3′ end of the MLK2 cDNA, a stretch of 13 adenine nucleotides is preceded by a typical polyadenylation signal (AATAAA) beginning 29 bases upstream from the poly(A) tract. This suggests that the 3′ end of the MLK2 mRNA is included in this insert.

Searches of nucleotide sequence databases with the complete MLK2 cDNA revealed the presence of four EST fragments with varying degrees of identity to MLK2. Only one of these fragments, from an infant brain cDNA library (GeneBank T15757), represents the 3′ terminus of the MLK2 message. The other three (GeneBank T98616, H01340, and H01390) all represent internal fragments of MLK2 cDNA. Both H01340 and H01390 are from a placental cDNA library and represent parts of the MLK2 SH3 and kinase domain sequence. T98616, however, is from a human foetal liver-spleen library and contains both a region of strong similarity to MLK2 and a region that is not similar to MLK2 at all. This EST may represent an artefact in the library or an alternately spliced product of the MLK2 gene.

Amino acid sequence. The longest open reading frame in the MLK2 cDNA encodes a putative protein of 954 amino acids with a calculated molecular mass of 103 506 Da. Based on a hydrophobicity analysis (Kyte and Doolittle, 1982) of the predicted amino acid sequence, the MLK2 protein contains no obvious signal sequence or membrane-spanning region (data not shown). Comparison of the sequence with conserved motifs for defined structural domains, however, reveals that MLK2 contains a SH3 domain, a kinase catalytic domain, and the unique double leucine zipper and basic domain of the MLK family (Dorow et al., 1993). The extreme N-terminal sequence of the MLK2 protein is acidic with four glutamic acid residues immediately following the putative initiator methionine (Fig. 1). This 22-amino-acid N-terminal sequence is unique with no significant similarity to any sequence in the protein data bases. The N-terminal region is followed by a 55-amino-acid sequence (residues 23−76) containing the highly conserved consensus motifs for SH3 domains (Musacchio et al., 1992). Alignment of this sequence with those of SH3 domains from the C-terminal of GRB2, phospholipase Cγ (PLC), spectrin (SPC), cellular *src* (*c-src*), neuronal *src* (*n-src*), and the p85 subunit of phosphatidylinositol 3′-kinase (p85-PtdIns3K) is shown in Fig. 2. The alignment shows that the MLK2-SH3 domain has a 5-amino-acid insert (residues 49−53) in a region corresponding to inserts of 6 and 15 amino acids in the SH3 domains of *n-src* (Martinez et al., 1987) and p85-PtdIns3K (Skolnick et al., 1991), respectively. These inserts are located in a region of the sequence that has been postulated to influence selectivity of SH3 domain binding (Booker et al., 1993; Koyama et al., 1993).

The MLK2 kinase catalytic domain, residues 101−359, contains all of the conserved amino acids forming the 11 sub-domain motifs used by Hanks et al. (1988) to define protein kinase families. Overall, the kinase domain amino acid sequence is more similar to that of the tyrosine than the serine/threonine-specific kinases. Furthermore, there are several motifs in the C-terminal portion of the MLK2 catalytic domain sequence that are found only in the tyrosine kinases and the *raf* oncogene family (Hanks et al., 1988). In particular, four motifs involving either tryptophan or proline residues are highly conserved in the receptor tyrosine kinases and the MLK family. In the MLK2 sequence these correspond to Trp294-Glu295, Cys339-Trp340, Pro301-Tyr302, and Cys328-Glu330. The MLK2 sequence, however, also contains a lysine residue (position 224) in subdomain VIb (Hanks et al., 1988) that is conserved in the serine/threonine, rather than the tyrosine, kinases. Thus, the MLK2 catalytic domain sequence falls into a category closely related to the tyrosine kinases but with predicted specificity for serine and/or threonine.

Beginning at residue 384 in the MLK2 amino acid sequence, there is an 80-amino-acid region that contains two leucine/iso-

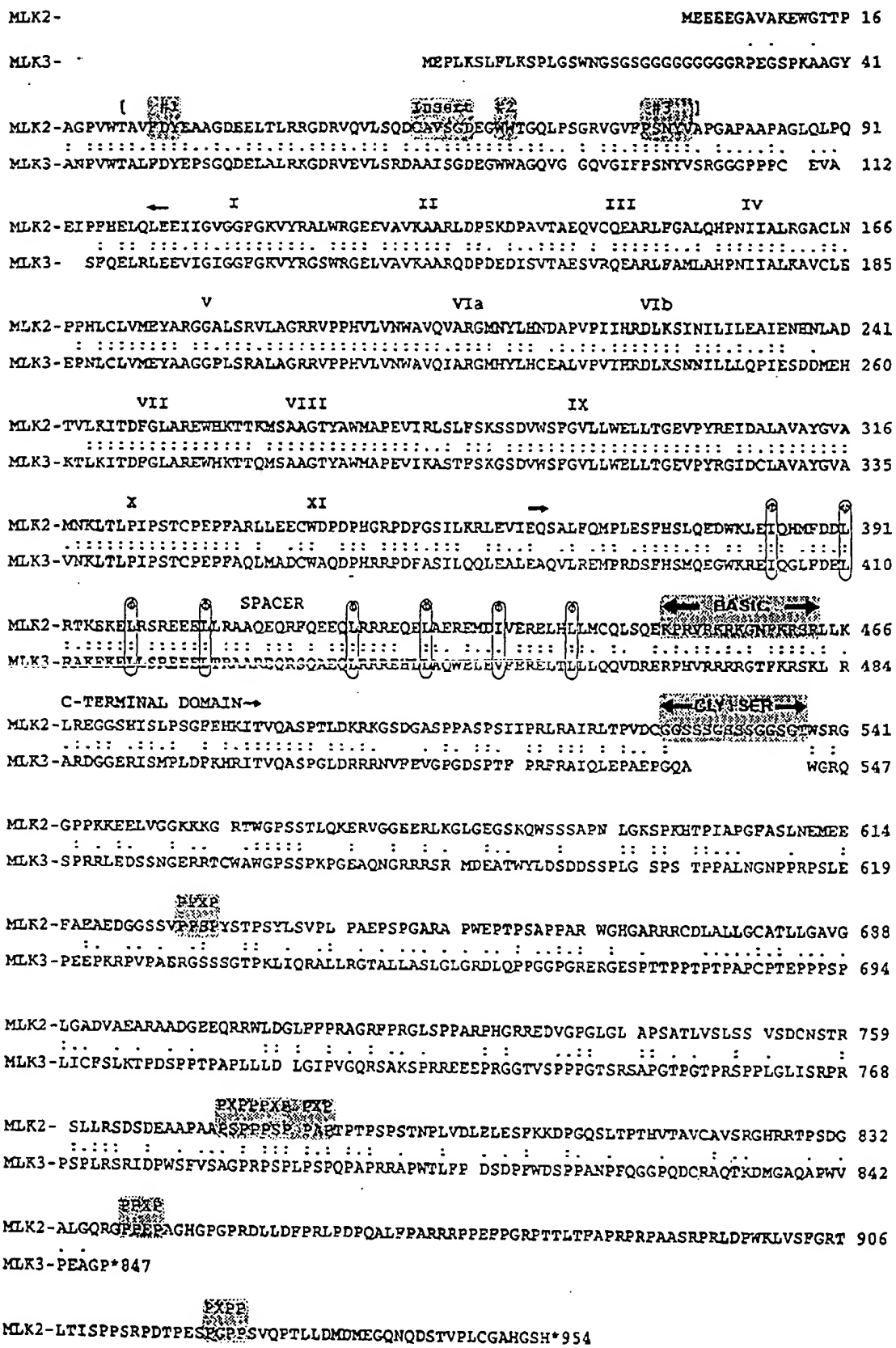


Fig. 3. Alignment of the predicted amino acid sequences of MLK2 and 3. Predicted amino acid sequences of MLK2 and 3 were aligned using the program CLUSTAL (Higgins and Sharp, 1988) and by eye. Protein domains are identified above the line. The SH3 domain is delineated by brackets and three SH3 consensus motifs containing aromatic residues are numbered within the brackets. The kinase domain is marked at the beginning and end by arrows and the subdomain motifs are numbered with roman numerals. Leucine zipper heptad residues are marked (*P*) and the beginning of the C-terminal domain is defined. The basic domain is marked at the beginning and end (⟷) as is the glycine/serine-rich peptide (⟷). Sequences similar to the core motif for SH3 domain recognition are marked with the letters P and X, where P represents a conserved proline and X any amino acid (shaded).

leucine heptad repeats and a basic motif. Each of the heptad motifs fits all of the criteria set out by Landschultz et al. (1988) for the classic leucine zippers of the transcription factors. Thus, they each contain 22 amino acids with a leucine or isoleucine residue at every seventh position, a higher than average content of charged amino acids (>50%), and an absence of the helix-breaking residues proline and glycine. The two zipper motifs are separated by a 13-amino-acid spacer sequence. A stretch of basic amino acids begins 9 residues after the last heptad leucine of the second zipper motif. Of 15 residues in this region, nine are basic (Lys or Arg) with no acidic residues. Furthermore, within this basic sequence is a motif (VRKRKG) that is very similar to nuclear localisation signals reported for several proteins, including the simian virus 40 large T antigen (reviewed by Kalderon et al., 1984).

Following the basic region, there is a C-terminal domain of 491 amino acids that is rich in serine/threonine (17%) and proline (16%). The high proline content is most striking in a stretch of 218 amino acids near the C-terminus (residues 712–929), where 22% of the residues are proline. One particular 20-amino-acid sequence (774–793) contains 11 proline, 4 serine, and 3 threonine residues. Furthermore, there are several poly(proline) motifs (Fig. 3) within the C-terminal domain that are similar to consensus sequences identified as binding sites for SH3 domains from a number of proteins (Ren et al., 1993; Yu et al., 1994). The amino acid composition of this domain is also biased toward basic, rather than acidic, residues with a calculated pI of 9.38.

Alignment of the predicted amino acid sequences of MLK2 and MLK3 (Fig. 3) shows a high degree of identity within their SH3, kinase catalytic, and leucine zipper/basic domains. This is most obvious for the kinase catalytic domains that share 83% amino acid sequence identity. If conservative substitutions are considered, the catalytic domain similarity between MLK2 and 3 is 90%. While the sequences of the SH3 domains have slightly reduced identity (70%), they share a very high degree of conservation of amino acids. Only 3 of 16 substitutions within the 55-amino-acid domain are non-conservative, corresponding to a similarity of 95%. There is, however, a single amino acid insertion in this domain in MLK2 (Ser65) compared to MLK3. There are several insertions/deletions between the two sequences in the non-conserved region joining the SH3 domain to the kinase catalytic domain (MLK2, residues 86–93). The insertion at position 65, however, is the only one that affects alignment of the MLK2 and 3 sequences within one of the known structural domains. The double leucine zipper and basic domain is also well conserved, with 65% identity/75% similarity between the two sequences.

Outside of the SH3, kinase and leucine zipper/basic domains, the similarity between MLK2 and MLK3 decreases dramatically. In the region closest to their N-termini, for instance, there is no similarity between the sequences. This N-terminal segment is 22 residues long in MLK2 and consists of 47 residues in MLK3. The four glutamic acid residues at the N-terminus of MLK2 are not present in the MLK3 sequence. The N-terminal region of MLK3, however, contains a 13-residue sequence with 11 glycine and 2 serine residues, including a stretch of nine consecutive glycine residues. It is interesting that this MLK3 poly(glycine) sequence is located at about the same distance from the beginning of the SH3 domain as the poly(glutamic acid) sequence in MLK2.

The large C-terminal domains of the two proteins, while both rich in serine/threonine and proline, are also poorly conserved. Aside from a few short sequences near the end of the basic domain, there is little actual identity between the two sequences. Furthermore, the size of the C-terminal domains in the two proteins differs quite considerably, with 491 amino acids in MLK2 compared to 360 in MLK3.

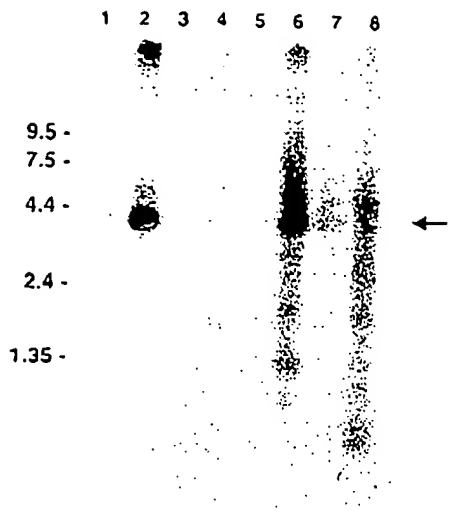


Fig. 4. Northern-blot analysis of human tissue mRNAs. Autoradiograph of an RNA blot hybridised with a MLK2 cDNA probe. Each lane contains 2 g mRNA from (1) heart, (2) brain, (3) placenta, (4) lung, (5) liver, (6) skeletal muscle, (7) kidney, and (8) pancreas. The MLK2 band is indicated (—). The blot was washed at a stringency of 0.5× NaCl/Cit at 65°C.

Northern-blot analysis. Expression of MLK2 RNA was examined by Northern-blot analysis of mRNA from human heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. The probe used for this analysis contained nucleotides 85–700 (Fig. 1) including 200 nucleotides of the 5′-untranslated MLK2 cDNA, as well as the sequence encoding the SH3 domain and the N-terminal region of the catalytic domain. In this analysis, a band at about 3.8 kb was detected at highest levels in RNA from brain and skeletal muscle, with a lower level in pancreas (Fig. 4). Expression in the other tissues was extremely low or undetectable. Given the high degree of sequence identity between MLK1–3, some cross-hybridisation in this type of assay might be expected. MLK3, however, has a much wider pattern of expression in human tissues, with high levels in lung, liver, and kidney as well as brain and skeletal muscle (Ing et al., 1994; Ezoe et al., 1994; Gallo et al., 1994). As hybridisation with the MLK2 probe did not reveal a significant level of expression in lung, liver or kidney, cross-hybridisation with MLK3 is not indicated. A MLK1 cDNA probe, however, hybridises to a band of an entirely different size compared to MLK2 (Dorow, D., unpublished results) ruling out possible cross-hybridisation with MLK1.

Chromosomal localisation. To determine the chromosomal location of the human MLK2 gene, metaphase chromosomes from two normal male donors were hybridised by the FISH technique using the 3454-bp MLK2 cDNA as a probe. 25 metaphases from one donor were examined, and a fluorescent signal was detected on one or both chromatids of all chromosomes 19. Of this signal, 73% was located at q13.2, with the remainder in the region of q13.1 to q13.3. There were 23 non-specific background signals recorded in the 25 metaphases. A similar result was obtained with the chromosome preparation from the second donor, confirming the localisation of the MLK2 gene to chromosome 19 q13.2.

DISCUSSION

We have reported the cDNA sequence, expression, and chromosomal location of MLK2, a member of the mixed-lineage family of protein kinases. To date, four members of this family have been described and complete cDNA sequences have now been reported for three, MLK2 (this study), MLK3 (Ing et al., 1994), and DLK (Holzman et al., 1994). MLK2 and 3 are each comprised of a SH3 domain, a kinase catalytic domain that is related to both the serine/threonine and the tyrosine types, a double leucine zipper, and a basic domain. In addition, they each have a large C-terminal domain rich in serine/threonine and proline. Our recent sequence data show that MLK1 also has a SH3 domain in its N-terminal region (Dorow, D., unpublished results). The fourth MLK family member, DLK (Holzman et al., 1994), lacks a SH3 domain, but otherwise shares significant structural similarity to MLK1–3.

SH3 domain. Among SH3 domains of known mammalian proteins, the SH3 domain sequences of the MLK proteins are most similar to that of the C-terminal SH3 domain of GRB2, an adaptor protein comprised of a SH2 domain flanked by SH3 domains (Lowenstein et al., 1992). GRB2 binds activated EGF receptors and platelet-derived-growth-factor receptors and connects them to the *ras* signalling pathway. The insert region, common to the SH3 domains of the MLK proteins, p85-PtdIns3K and n-*src*, however, is not found in the GRB2-SH3 domain sequence.

Three-dimensional structures of SH3 domains from several proteins, including α-spectrin (Musacchio et al., 1992), c-Fyn (Noble et al., 1993), Lck (Eck et al., 1994), c-*src* (Yu et al., 1992), p85-PtdIns3K (Booker et al., 1993; Koyama et al., 1993), and phospholipase Cγ (Khoda et al., 1993) have been published. While SH3 domains from the various proteins have a low degree of amino acid identity, all show a similar folding pattern. There are a series of motifs with several conserved aromatic residues, corresponding to MLK2 residues 24–26 (Phe-Asp-Tyr), 57 and 58 (Trp-Trp), and 71–75 (Pro-Ser-Asn-Tyr-Val), that contribute to the SH3 domain consensus (Koch et al., 1991). In the folded structures, the conserved residues are located in β-sheets connected by variable loops and the aromatic side chains line a binding pocket on the surface of the protein (Koyama et al., 1993; Booker et al., 1993; Yu et al., 1994). Part of one variable loop forms an end of the binding pocket, leading to speculation that residues within this loop may contribute to fine specificity of the domain. The inserts in the SH3 domains of MLK2, MLK3, p85-PtdIns3K, and n-*src* are all located within this loop. The placement of the inserts at one end of the binding pocket suggests that they may play a role in target recognition by these SH3 domains. This is further supported by the differential peptide recognition of the SH3 domains of c-*src* and n-*src* (Cicchetti et al., 1992; Booker et al., 1993), that are almost identical except for the presence of the n-*src* insert. The only other difference between the two being a Ser → Thr substitution in another variable loop (Fig. 2). While the 5-residue insert in the MLK family contains only hydrophobic and polar residues, four of the six residues within the n-*src* insert are charged. In the longer p85-PtdIns3K insert, the first six residues are hydrophobic with several charged residues in the remaining nine. These differences suggest that the insert sequence may be an important determinant of specificity of the MLK-SH3 domain.

A second variable area, corresponding to residues 63–67 of MLK2, is located between the last two consensus motifs of the SH3 domain sequence. Within this region there are three replacements and one insertion in MLK2, compared to MLK3. The sequence of this loop is Leu-Pro-Ser-Gly-Arg in MLK2 compared to Val-Gly-Gly-Gln in MLK3. Sequence variation within this region, therefore, may also effect recognition specificity of MLK-SH3 domains.

Catalytic domain. As has been suggested for other kinase family members (Hanks et al., 1988), the high degree of catalytic domain identity between the MLK proteins implies that they may also have similarity in their cellular roles. The kinase domain sequences of MLK1–3 are highly conserved and closely related to members of the tyrosine-specific kinase families (Hanks, 1991). The kinase domain of DLK, while conserving the general features of the MLK enzymes, shares only 36% amino acid identity to MLK1 (Holzman et al., 1994) and several gaps must be introduced to align the sequences. DLK, however, has very strong similarity to a 108-residue fragment of a putative human serine/threonine kinase in the PIR data base (accession number S37420; Schultz and Nigg, 1993). Within the region of overlap, these two sequences share 87% identity and 92% similarity (Holzman et al., 1994). As has been suggested by Holzman et al. (1994), DLK and the protein represented by PIR accession number S37420, appear to form a second more distantly related subgroup within the MLK family. Both DLK and MLK3 (SPRK) autophosphorylate on serine and threonine in immune complex kinase assays (Holzman et al., 1994; Gallo et al., 1994). To date, no activity of any of the MLK proteins toward other substrates has been demonstrated. It is therefore not possible to rule out specificity for tyrosine phosphorylation of some target protein. It is most probable, however, that the MLK enzymes will all display serine/threonine activity. In this respect, SPRK is the only kinase with demonstrated serine/threonine specificity to also have a SH3 domain (Gallo et al., 1994).

The double leucine zipper domain. The double leucine zipper domain is the most striking structural feature of the MLK sequences. The overall structure of this domain is highly conserved in all members of the MLK family. Furthermore, the distance between the kinase domain and the first leucine of zipper motif no. 1 is almost identical between MLK1–3 and DLK. In DLK, however, there is an 18-amino-acid insert in the spacer region between the two zipper sequences. Thus, the zipper motifs are 13 residues apart in MLK1–3 and 31 in DLK. As leucine zippers are commonly found in proteins that form dimers or higher order oligomers, the zippers of MLK proteins may participate in formation of complexes with themselves, other members of the family, or other proteins. Chou and Fasman (1978) analysis of the double zipper region of MLK1–3 has predicted a helix-turn-helix conformation for this domain (Dorow et al., 1993; Gallo et al., 1994). As discussed previously (Dorow et al., 1993), such a conformation would allow for interaction between the two zippers of one MLK molecule to form a uniquely folded zipper-turn-zipper domain.

The conserved basic sequence following the second zipper motif in MLK1–3 is placed at about the same distance from the leucine zipper motifs as the basic DNA-binding sequence of the transcription factors. In the transcription factors, however, the DNA-binding peptide is on the N-terminal side of the leucine zipper. The basic sequence is not conserved in the DLK sequence, although DLK contains one similarly placed stretch of eight amino acids of which four are basic (Holzman et al., 1994). The basic domains of MLK1–3 all contain a motif that is consistent with nuclear localisation signals that function in several proteins (Kalderon et al., 1984). It has not yet been determined, however, if the basic sequence functions as a nuclear targeting signal in any of the MLK proteins.

C-terminal serine/threonine and proline-rich domain. While the C-terminal domain amino acid sequences of MLK2 and 3

each comprises a similar high serine/threonine and proline content, the actual sequences have little identity. There are a few short segments that share sequence similarity, but these constitute only a minute proportion of the total sequences of the C-terminal domains. DLK contains a 332-amino-acid C-terminal domain that is also rich in serine and proline. The C-terminal domain was not coded by the human colonic cDNA clone from which the MLK2 partial amino acid sequence was previously predicted (Dorow et al., 1993). In the cDNA sequence of the original clone, there was a one-base insertion causing a shift to a reading frame with a premature stop codon. This appears to have been an artefact in that particular clone, as another clone subsequently isolated from a human colonic cDNA library matches that reported here from the human brain.

Due to the high content of hydroxy amino acids, the C-terminal domain of MLK2 may be a target for regulatory phosphorylation events. There is one 13-residue segment (residues 524–537) comprised of six glycine, seven serine, and one threonine residue, that is not conserved in the MLK3 C-terminal sequence. This segment, however, does have some similarity to the glycine rich sequence in the N-terminal region of MLK3. Within the MLK2 C-terminal domain there are also several proline-rich motifs that are similar, but not identical, to sequences identified as binding peptides for SH3 domains from several proteins (Ren et al., 1993; Yu et al., 1994). Recently, two models have been put forward to explain the mechanism of SH3 domain recognition of poly(proline) peptides (Feng et al., 1994; Lim and Richards, 1994). Both of these models suggest that the minimal sequence required for SH3 domain recognition is Pro-Xaa-Xaa-Pro. In an extensive study, Feng et al. (1994) used both sequence comparisons and mutagenic analysis to propose a general scheme for recognition. In this analysis two classes of binding peptides were identified. The peptides conformed to either Xaa-p-Xaa-P-p-Xaa (class I) and Xaa-P-p-Xaa-P-p-Xaa (class II) consensus sequences (where Xaa is any amino acid, p is a scaffolding residue that is often proline and P is a critical proline residue). Within the C-terminal domain of MLK2 there are four P-P-Xaa-P and three P-Xaa-P-P sequences. There is also one particular proline-rich motif with the sequence PSPPPSPPAPTPTP that contains both the class I and class II consensus sequences. These proline-rich sequences present the possibility for an intramolecular interaction between the MLK2-SH3 domain and its own C-terminal domain. This same possibility has been noted for MLK3 (Ing et al., 1994).

Tissue distribution. By Northern-blot analysis of mRNAs from several human tissues, MLK2 expression was highest in brain and skeletal muscle, with a reduced signal in pancreas. MLK2 is also expressed at low levels in epithelial cell lines of the breast and colon, but was not detected in haemopoietic cell lines (Dorow, D., unpublished results). MLK3, however, has a wide expression pattern with moderate expression in most cell lines and tissues examined (Ing et al., 1994; Gallo et al., 1994; Ezoe et al., 1994). Expression of MLK3 is high in placenta, lung, and liver whereas the expression of MLK2 is barely detectable. MLK3 expression has also been found to be high in melanocytes and in cells of haematopoietic origin (Ing et al., 1994; Ezoe et al., 1994). The expression pattern of murine DLK is more restricted, with significant expression being detected only in foetal and adult brain (Holzman et al., 1994).

Chromosomal localisation. The gene encoding MLK2 has been mapped to human chromosome 19 q13.1–13.3. The distribution of signal in this analysis suggests that the gene resides in the q13.2 region. A search of the Genome Data Base (William H. Welch Medical Library, Johns Hopkins University, Baltimore, USA) revealed that other genes mapped to human 19 q13.2 include apolipoproteins C-I, C-II, and E, several carcinoembryonic antigen and pregnancy-specific β-1 glycoprotein family members, and transforming growth factor β-1. The gene for MLK-1 has been localised to human chromosome 14 q24.3–31 (Dorow et al., 1993) while that of MLK3 was mapped to 11 q13.1–13.3 (Ing et al., 1994).

Thus far, the role of MLK proteins in cellular networks has not been elucidated. The presence within their sequences of domains associated with regulation of signal transduction, phosphorylation and gene transcription, however, suggests that they may play important roles in the control of cellular processes. Identification of proteins with which the different MLK2 domains interact may provide information on new pathways for cellular regulation.

The authors wish to thank Dr Simon Stuart, Prof. Claude Bernard, and Dr Ora Bernard for the gift of the human brain cDNA library; the authors also thank James Eddes for help with the preparation of figures. This work was supported by grants from the National Health and Medical Research Council of Australia (D. S. D. and R. J. S.; J. K. N. and G. R. S.), the J. H. & J. D. Gunn Medical Research Foundation (J. K. N. and G. R. S.), and a Howard Hughes Medical Institute International Research Scholars Award to G. R. S.

REFERENCES

Bar-Sagi, D., Rotin, D., Batzer, A., Mandiyan, V. & Schlessinger, J. (1993) SH3 domains direct cellular localization of signalling molecules, *Cell 74*, 83–91.

Birchmeier, C., Sonnenberg, E., Weidner, K. M. & Walter, B. (1993) Tyrosine kinase receptors in the control of epithelial growth and morphogenesis during development, *Bioessays 15*, 185–189.

Booker, G. W., Gout, I., Downing, A. K., Driscoll, P. C., Boyd, J., Waterfield, M. D. & Campbell, I. D. (1993) Solution structure and ligand-binding site of the SH3 domain of the p85α subunit of phosphatidylinositol 3-kinase, *Cell 73*, 813–822.

Callen, D. F., Baker, E., Eyre, H. J., Chernos, J. E., Bell, J. A. & Sutherland, G. R. (1990) Reassessment of two apparent deletions of chromosome 16p to an ins (11;16) and at (1;16) by chromosome painting, *Ann. Génét. 33*, 219–221.

Chou, P. Y. & Fasman, G. D. (1978) Empirical predictions of protein conformation, *Annu. Rev. Biochem. 47*, 251–276.

Cicchetti, P., Mayer, B. J., Thiel, G. & Baltimore, D. (1992) Identification of a protein that binds to the SH3 region of abl and is similar to bcr and GAP-rho, *Science 257*, 803–806.

Dorow, D. S., Devereux, L., Dietzsch, E. & deKretser, T. (1993) Identification of a new family of human epithelial protein kinases containing two leucine/isoleucine zipper domains, *Eur. J. Biochem. 231*, 701–710.

Dorow, D. S., Devereux, L. & Simpson, R. J. (1994) The mixed lineage kinases: a family of protein kinases containing a double leucine zipper domain, a basic motif and SH3 domain, *J. Protein Chem. 13*, 458–460.

Drubin, D. G., Mulholland, J., Zhu, Z. & Botstein, D. (1990) Homology of a yeast actin-binding protein to signal transduction proteins and myosin-1, *Nature 343*, 288–290.

D'Urso, G., Marraccino, R. L., Marshak, D. R. & Roberts, J. M. (1990) Cell cycle control of DNA replication by a homologue from human cells of the p34^{cdc2} protein kinase, *Science 250*, 786–791.

Eck, M. J., Atwell, S. K., Shoelson, S. E. & Harrison, S. C. (1994) Structure of the regulatory domains of the Src-family tyrosine kinases Lck, *Nature 368*, 764–769.

Ezoe, K., Lee, S.-T., Strunk, K. & Spritz, R. A. (1994) PTK1, a novel protein kinase required for proliferation of human melanocytes, *Oncogene 9*, 935–938.

Feinberg, A. P. & Vogelstein, B. (1983) A technique for labelling restriction endonuclease fragments to high specific activity, *Anal. Biochem. 132*, 6–13.

Feng, S., Chen, J. K., Yu, H., Simon, J. A. & Schreiber, S. L. (1994) Two binding orientations for peptides to the src SH3 domain: devel-

opment of a general model for SH3-ligand interactions. *Science 266*, 1241–1247.

Gallo, K. A., Mark, M. R., Scadden, D. T., Wang, Z., Gu, Q. & Godwoski, P. J. (1994) Identification and characterization of SPRK, a novel src-homology 3 domain-containing proline-rich kinase with serine/threonine kinase activity, *J. Biol. Chem. 269*, 15092–15100.

Goblet, C., Prost, E. & Whalen, R. G. (1989) One-step amplification of transcripts in total RNA using polymerase chain reaction. *Nucleic Acids Res. 17*, 2144.

Gout, I., Dhand, R., Hiles, I. D., Fry, M. J., Panayotou, G., Das, P., Truong, O., Totty, N. F., Hsuan, J., Booker, G. W., Campbell, I. D. & Waterfield, M. D. (1993) The GTPase dynamin binds to and is activated by a subset of SH3 domains. *Cell 75*, 25–36.

Hanks, S. K., Quinn, A. M. & Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science 241*, 42–52.

Hanks, S. K. (1991) Eukaryotic protein kinases. *Curr. Opin. Struct. Biol. 1*, 369–383.

Higgins, D. G. & Sharp, P. M. (1988) Clustal: a package for performing multiple sequence alignments on a microcomputer. *Gene (Amst.) 73*, 237–244.

Holzman, L. B., Merritt, S. E. & Fan, G. (1994) Identification, molecular cloning and characterization of dual leucine zipper bearing kinase. *J. Biol. Chem. 269*, 30808–30817.

Hunter, T. & Karin, M. (1992) The regulation of transcription by phosphorylation. *Cell 70*, 375–387.

Hunter, T. & Pines, J. (1994) Cyclins and cancer II: cyclin D and CDK inhibitors come of age, *Cell 79*, 573–582.

Ing, Y. L., Leung, I. W. L., Heng, H. H. Q., Tsui, L.-C. & Lassam, N. J. (1994) MLK-3: identification of a widely-expressed protein kinase bearing an SH3 and leucine zipper-basic region domain. *Oncogene 9*, 1745–1750.

Kalderon, D., Richardson, W. D., Markham, A. T. & Smith, A. E. (1984) Sequence requirements for nuclear localization of simian virus large-T antigen, *Nature 311*, 33–38.

Khoda, D., Hatanaka, H., Odaka, M., Mandiyan, V., Ullrich, A., Schlessinger, J. & Inagaki, F. (1993) Solution structure of the SH3 domain of phospholipase C-γ, *Cell 72*, 953–960.

Koch, C. A., Anderson, D., Moran, M. F., Ellis, C. & Pawson, T. (1991) SH2 and SH3 domains: elements that control interactions of cytoplasmic signalling proteins. *Science 252*, 668–674.

Koyama, S., Yu, H., Dalgarno, D. C., Shin, T. B., Zydowsky, L. D. & Schreiber, S. L. (1993) Structure of the PI3K SH3 domain and analysis of the SH3 family. *Cell 72*, 945–952.

Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation of eukaryotic ribosomes. *Cell 44*, 283–292.

Kozak, M. (1989) The scanning model for translation: an update. *J. Cell. Biol. 108*, 229–241.

Kyte, J. & Doolittle, R. F. (1982) A simple method for displaying the hydrophobic nature of a protein. *J. Mol. Biol. 157*, 105–132.

Landschultz, W. H., Johnson, P. F. & McKnight, S. L. (1988) The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science 240*, 1759–1764.

Lim, W. A. & Richards, F. M. (1994) Critical residues in an SH3 domain from sem-5 suggest a mechanism for proline-rich peptide recognition. *Struct. Biol. 1*, 221–225.

Lowenstein, E. J., Daly, R. J., Batzer, A. G., Li, W., Margolis, B., Lammers, R., Ullrich, A. & Schlessinger, J. (1992) The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signalling, *Cell 70*, 431–442.

Martinez, R., Mathey-Prevot, B., Bernards, A. & Baltimore, D. (1987) Neuronal pp60$^{c\text{-}src}$ contains a six-amino acid insertion relative to its non-neuronal counterpart. *Science 237*, 411–415.

Moore, G. E., Merrick, S. B., Woods, L. K. & Arabasz, N. M. (1975). A human squamous cell carcinoma cell line. *Cancer Res. 35*, 2684–2688.

Musacchio, A., Noble, M., Pauptit, R., Wierenga, R. & Saraste, M. (1992) Crystal structure of a Src-homology 3 (SH3) domain, *Nature 359*, 851–855.

Noble, M. E. M., Musacchio, A., Saraste, M., Courtneidge, S. A. & Wierenga, R. K. (1993) Crystal structure of the SH3 domain in human Fyn: comparison of the three-dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J. 12*, 2617–2624.

O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991) X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science 254*, 539–544.

Pawson, T. & Gish, G. D. (1992) SH2 and SH3 domains: from structure to function, *Cell 71*, 359–362.

Posada, J. & Cooper, J. A. (1992) Molecular signal integration. Interplay between serine, threonine, and tyrosine phosphorylation. *Mol. Biol. Cell 3*, 583–592.

Ren, R., Mayer, B. J., Cicchetti, P. & Baltimore, D. (1993) Identification of a ten-amino acid proline-rich SH3 binding site, *Science 259*, 1157–1161.

Rodaway, A. R. F., Sternberg, M. J. E. & Bentley, D. L. (1989) Similarity in membrane proteins. *Nature 342*, 624.

Sadowski, F., Stone, J. C. & Pawson, T. (1986) A non catalytic domain conserved among cytoplasmic protein tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus p130gag-fps, *Mol. Cell. Biol. 6*, 4396–4408.

Schultz, S. J. & Nigg, E. A. (1993) Identification of 21 novel human protein kinases, including 3 members of a family related to the cell cycle regulator nimA for *Aspergillus nidulans*. *Cell Growth & Differ. 4*, 821–830.

Skolnick, E. Y., Margolis, B., Mohammadi, M., Lowenstein E., Fisher, R., Drepps, A., Ullrich, A. & Schlessinger, Y. (1991) Cloning of PI3 kinase-associated p85 utilizing a novel method for expression/cloning of target proteins for receptor tyrosine kinases. *Cell 65*, 83–90.

Superti-Furga, G., Fumagalli, S., Koegl, M., Courtneidge, S. A. & Draetta, G. (1993) Csk inhibition of c-Src activity requires both the SH2 and SH3 domains of Src, *EMBO J. 12*, 2625–2634.

Vinson, C. R., Sigler, P. B. & McKnight, S. L. (1989) Scissors-grip model for DNA recognition by a family of leucine zipper proteins, *Science 246*, 911–916.

Wheater, R. F. & Roberts, S. H. (1987) An improved lymphocyte culture technique: deoxycytidine release of a thymidine block and use of a constant humidity chamber for slide making. *J Med. Genet. 24*, 113–115.

Wilks, A. F. (1989) Two putative protein-tyrosine kinases identified by application of the polymerase chain reaction. *Proc. Natl Acad. Sci. USA 86*, 1603–1607.

Wolfe, L., Corbin, J. D. & Francis, S. H. (1989) Characterization of a novel isozyme of cGMP dependent kinase from bovine aorta, *J. Biol. Chem. 264*, 7734–7741.

Yu, H., Rosen, M., Shin, T. B., Seidell-Dugan, C., Brugge, J. S. & Schreiber, S. L. (1992) Solution structure of the SH3 domain of src and identification of its ligand binding site. *Science 258*, 1665–1668.

Yu, H., Chen, J. K., Feng, S., Dalgarno, D. C., Brauer, A. W. & Schreiber, S. L. (1994) Structure basis of binding of proline-rich peptides to SH3 domains. *Cell 76*, 933–945.

Appendix B

Cell, Vol. 72, 971–983, March 26, 1993, Copyright © 1993 by Cell Press

A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Summary

The Huntington's disease (HD) gene has been mapped to 4p16.3 but has eluded identification. We have used haplotype analysis of linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the defect. A new gene, IT15, isolated using cloned trapped exons from the target area contains a polymorphic trinucleotide repeat that is expanded and unstable on *HD* chromosomes. A $(CAG)_n$ repeat longer than the normal range was observed on *HD* chromosomes from all 75 disease families examined, comprising a variety of ethnic backgrounds and 4p16.3 haplotypes. The $(CAG)_n$ repeat appears to be located within the coding sequence of a predicted ~348 kd protein that is widely expressed but unrelated to any known gene. Thus, the *HD* mutation involves an unstable DNA segment, similar to those described in fragile X syndrome, spino-bulbar muscular atrophy, and myotonic dystrophy, acting in the context of a novel 4p16.3 gene to produce a dominant phenotype.

Introduction

Huntington's disease (HD) is a progressive neurodegenerative disorder characterized by motor disturbance, cognitive loss, and psychiatric manifestations (Martin and Gusella, 1986). It is inherited in an autosomal dominant fashion and affects ~1 in 10,000 individuals in most populations of European origin (Harper et al., 1991). The hallmark of HD is a distinctive choreic movement disorder that typically has a subtle, insidious onset in the fourth to fifth decade of life and gradually worsens over a course of 10 to 20 years until death. Occasionally, HD is expressed in juveniles, typically manifesting with more severe symptoms including rigidity and a more rapid course. Juvenile onset of HD is associated with a preponderance of paternal transmission of the disease allele. The neuropathology of HD also displays a distinctive pattern, with selective loss of neurons that is most severe in the caudate and putamen. The biochemical basis for neuronal death in HD has not yet been explained, and there is consequently no treatment effective in delaying or preventing the onset and progression of this devastating disorder.

The genetic defect causing HD was assigned to chromosome 4 in 1983 in one of the first successful linkage analyses using polymorphic DNA markers in humans (Gusella

*The Huntington's Disease Collaborative Research Group comprises:

Group 1:
Marcy E. MacDonald,[1] Christine M. Ambrose,[1] Mabel P. Duyao,[1] Richard H. Myers,[2] Carol Lin,[1] Lakshmi Srinidhi,[1] Glenn Barnes,[1] Sherryl A. Taylor,[1] Marianne James,[1] Nicolet Groot,[1] Heather MacFarlane,[1] Barbara Jenkins,[1] Mary Anne Anderson,[1] Nancy S. Wexler,[3] and James F. Gusella[1†]
[1]Molecular Neurogenetics Unit
Massachusetts General Hospital
and Department of Genetics
Harvard Medical School
Boston, Massachusetts 02114
[2]Department of Neurology
Boston University Medical School
Boston, Massachusetts 02118
[3]Hereditary Disease Foundation
1427 7th Street, Suite 2
Santa Monica, California 90401

Group 2:
Gillian P. Bates, Sarah Baxendale, Holger Hummerich, Susan Kirby, Mike North, Sandra Youngman, Richard Mott, Gunther Zehetner, Zdenek Sedlacek, Annemarie Poustka, Anna-Marie Frischauf, and Hans Lehrach
Genome Analysis Laboratory
Imperial Cancer Research Fund
Lincoln's Inn Fields
London, WC2A 3PX, England

Group 3:
Alan J. Buckler,[1] Deanna Church,[1] Lynn Doucette-Stamm,[1] Michael C. O'Donovan,[1] Laura Riba-Ramirez,[1] Manish Shah,[1] Vincent P. Stanton,[1] Scott A. Strobel,[2] Karen M. Draths,[2] Jennifer L. Wales,[2] Peter Dervan,[2] and David E. Housman[1]
[1]Center for Cancer Research
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
[2]Division of Chemistry and Chemical Engineering
California Institute of Technology
Pasadena, California 91125

Group 4:
Michael Altherr, Rita Shiang, Leslie Thompson, Thomas Fielder, and John J. Wasmuth
Department of Biological Chemistry
University of California
Irvine, California 92717

Group 5:
Danilo Tagle, John Valdes, Lawrence Elmer, Marc Allard, Lucio Castilla, Manju Swaroop, Kris Blanchard, and Francis S. Collins
Department of Internal Medicine and Human Genetics
and The Howard Hughes Medical Institute
University of Michigan
Ann Arbor, Michigan 48109

Group 6:
Russell Snell, Tracey Holloway, Kathleen Gillespie, Nicole Datson, Duncan Shaw, and Peter S. Harper
Institute of Medical Genetics
University of Wales College of Medicine
Cardiff, CF4 4XN, Wales

†*Correspondence should be addressed to James F. Gusella.*

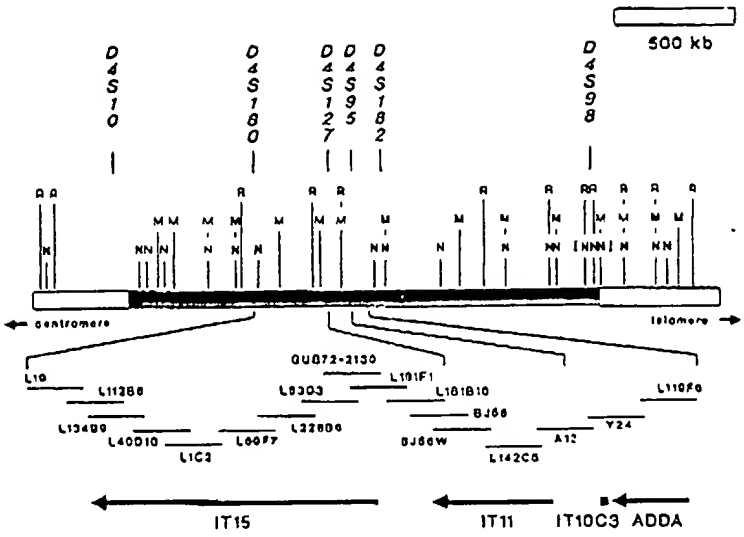


Figure 1. Long-Range Restriction Map of the *HD* Candidate Region

A partial long-range restriction map of 4p16.3 is shown (adapted from Lin et al. [1991]). The *HD* candidate region determined by recombination events is depicted by hatched bars between *D4S10* and *D4S98*. The portion of the *HD* candidate region implicated as the site of the defect by linkage disequilibrium haplotype analysis (MacDonald et al., 1992) is shown as a closed bar. Below the schematic map, the region from *D4S180* to *D4S182* is expanded to show the cosmid contig (averaging 40 kb per cosmid). The genomic coverage and, where known, the transcriptional orientation (arrows, 5' to 3') of the IT15, IT11, IT10C3, and *ADDA* genes is also shown. Locus names above the map denote selected polymorphic markers that have been used in HD families. The positions of *D4S127* and *D4S95*, which form the core of haplotype in the region of maximum disequilibrium, are also shown in the cosmid contig. Restriction sites are given for NotI (N), MluI (M), and NruI (R). Sites displaying complete digestion are shown in boldface, while sites subject to frequent incomplete digestion are shown as lighter symbols. Brackets around the N symbols indicate the presence of additional clustered NotI sites.

et al., 1983). Since that time, we have pursued a location cloning approach to isolating and characterizing the *HD* gene based on progressively refining its localization (Gusella, 1989, 1991). Among other work, this has involved the generation of new genetic markers in the region by a number of techniques (Pohl et al., 1988; Whaley et al., 1991; MacDonald et al., 1989a), the establishment of genetic (MacDonald et al., 1989b; Allitto et al., 1991) and physical maps of the implicated regions (Bucan et al., 1990; Bates et al., 1991; Doucette-Stamm et al., 1991; Altherr et al., 1992), the cloning of the 4p telomere of an *HD* chromosome in a yeast artificial chromosome clone (Bates et al., 1990; Youngman et al., 1992), the establishment of yeast artificial chromosome (Bates et al., 1992) and cosmid (S. B. et al., unpublished data) contigs of the candidate region, as well as the analysis and characterization of a number of candidate genes from the region (Thompson et al., 1991; Taylor et al., 1992; Ambrose et al., 1992; M. P. D. et al., submitted). Analysis of recombination events in HD kindreds has identified a candidate region of 2.2 Mb, between *D4S10* and *D4S98* in 4p16.3, as the most likely position of the *HD* gene (MacDonald et al., 1989b; Bates et al., 1991; Snell et al., 1992). Investigations of linkage disequilibrium between *HD* and DNA markers in 4p16.3 (Snell et al., 1989; Theilman et al., 1989) have suggested that multiple mutations have occurred to cause the disorder (MacDonald et al., 1991). However, haplotype analysis using multiallele markers has indicated that at least one-third of *HD* chromosomes are ancestrally related (MacDonald et al., 1992). The haplotype shared by these *HD* chromosomes indicates that a 500 kb segment between *D4S180* and *D4S182* is the most likely site of the genetic defect.

Targeting this 500 kb region for saturation with gene transcripts, we have used exon amplification as a rapid method for obtaining candidate coding sequences (Buckler et al., 1991). This strategy has previously identified three genes: the α-adducin gene (ADDA) (Taylor et al., 1992) and a putative novel transporter gene (IT10C3) in the distal portion of this segment (M. P. D. et al., submitted), and a novel G protein–coupled receptor kinase gene (IT11) in the central portion (Ambrose et al., 1992). However, no defects implicating any of these genes as the *HD* locus have been found. We have now applied the exon amplification approach to the proximal portion of the 500 kb segment. We have identified a large gene, IT15, spanning ~210 kb, that encodes a previously undescribed protein of ~348 kd. The IT15 reading frame contains a polymorphic (CAG)$_n$ trinucleotide repeat with at least 17 alleles in the normal population, varying from 11 to 34 CAG copies. On *HD* chromosomes, the length of the trinucleotide repeat is substantially increased, to a range of 42 to over 66 copies, and shows an apparent correlation with age of onset, the longest segments being detected in juvenile HD cases. The instability in the length of the repeat is reminiscent of similar trinucleotide repeats in the fragile X syndrome and in myotonic dystrophy (Suthers et al., 1992). The presence of an unstable, expandable trinucleotide repeat on *HD* chromosomes in the region of strongest linkage disequilibrium with the disorder suggests that this alteration underlies the dominant phenotype of *HD* and that IT15 encodes the *HD* gene.

Results

Application of Exon Amplification to Obtain Trapped, Cloned Exons

The *HD* candidate region defined by discrete recombina-

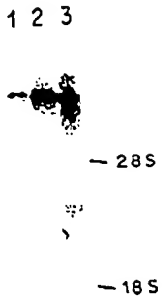


Figure 2. Northern Blot Analysis of the IT15 Transcript

Results of the hybridization of IT15A to a Northern blot of RNA from normal (lane 1) and *HD* homozygous (lanes 2 and 3) lymphoblasts are shown. A single RNA of ~11 kb was detected in all three samples, with slight apparent variations being due to unequal RNA concentrations. The *HD* homozygotes are independent, deriving from an American family (lane 2) and the large Venezuelan family (lane 3), respectively. The Venezuelan *HD* chromosome has a 4p16.3 haplotype of 5 2 2 defined by a (GT)$_n$ polymorphism at *D4S127* and variable number tandem repeat and TaqI restriction fragment length polymorphisms at *D4S95*. The American homozygote carries the most common 4p16.3 haplotype found on *HD* chromosomes: 2 11 1 (MacDonald et al., 1992).

tion events in well-characterized families spans 2.2 Mb between *D4S10* and *D4S98* as shown in Figure 1. The 500 kb segment between *D4S180* and *D4S182* displays the strongest linkage disequilibrium with *HD*, with about one-third of disease chromosomes sharing a common haplotype, anchored by multiallele polymorphisms at *D4S127* and *D4S95* (MacDonald et al., 1992). We have isolated 64 overlapping cosmids, spanning ~480 kb from *D4S180* to a location between *D4S95* and *D4S182*, by a combination of information from yeast artificial chromosome (Baxendale et al., 1991) and cosmid probe hybridization to high density filter grids of a chromosome 4–specific library, as well as of additional libraries covering this region. Sixteen of these cosmids providing the complete contig are shown in Figure 1. We have previously used exon amplification to identify *ADDA*, IT10C3 (a novel putative transporter gene), and IT11 (a novel G protein–coupled receptor kinase gene) in the region distal to *D4S127* (Figure 1).

We have now applied the exon amplification technique to cosmids from the region of the contig proximal to *D4S127*. This procedure produces trapped exon clones, which can represent single exons or multiple exons spliced together, and is an efficient method for obtaining probes for screening cDNA libraries. Individual cosmids were processed, yielding nine exon clones in the region from cosmids L13489 to L181B10.

Identification of the IT15 Gene

Two nonoverlapping cDNAs were initially isolated using exon probes. IT15A was obtained by screening a transformed adult retinal cell cDNA library with exon clone DL118F5-U. IT16A was isolated by screening an adult frontal cortex cDNA library with a pool of three exon clones, DL83D3-8, DL83D3-1, and DL228B6-3. By Northern blot analysis, we discovered that IT15A and IT16A both detected an ~10–11 kb transcript, suggesting that they derive from the same mRNA. Figure 2 shows an example of a Northern blot containing RNA from lymphoblastoid cell lines representing a normal individual and two independent homozygotes for *HD* chromosomes of different haplotypes. The same ~10–11 kb transcript was also detected in RNA from a variety of human tissues (liver, spleen, kidney, muscle, and various regions of adult brain).

IT15A and IT16A were used to "walk" in a number of human tissue cDNA libraries in order to obtain the full-length transcript. Figure 3 shows a representation of five cDNA clones that define the IT15 transcript, under a sche-

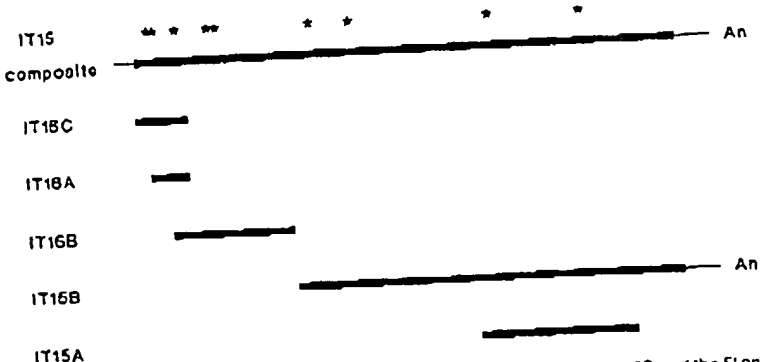
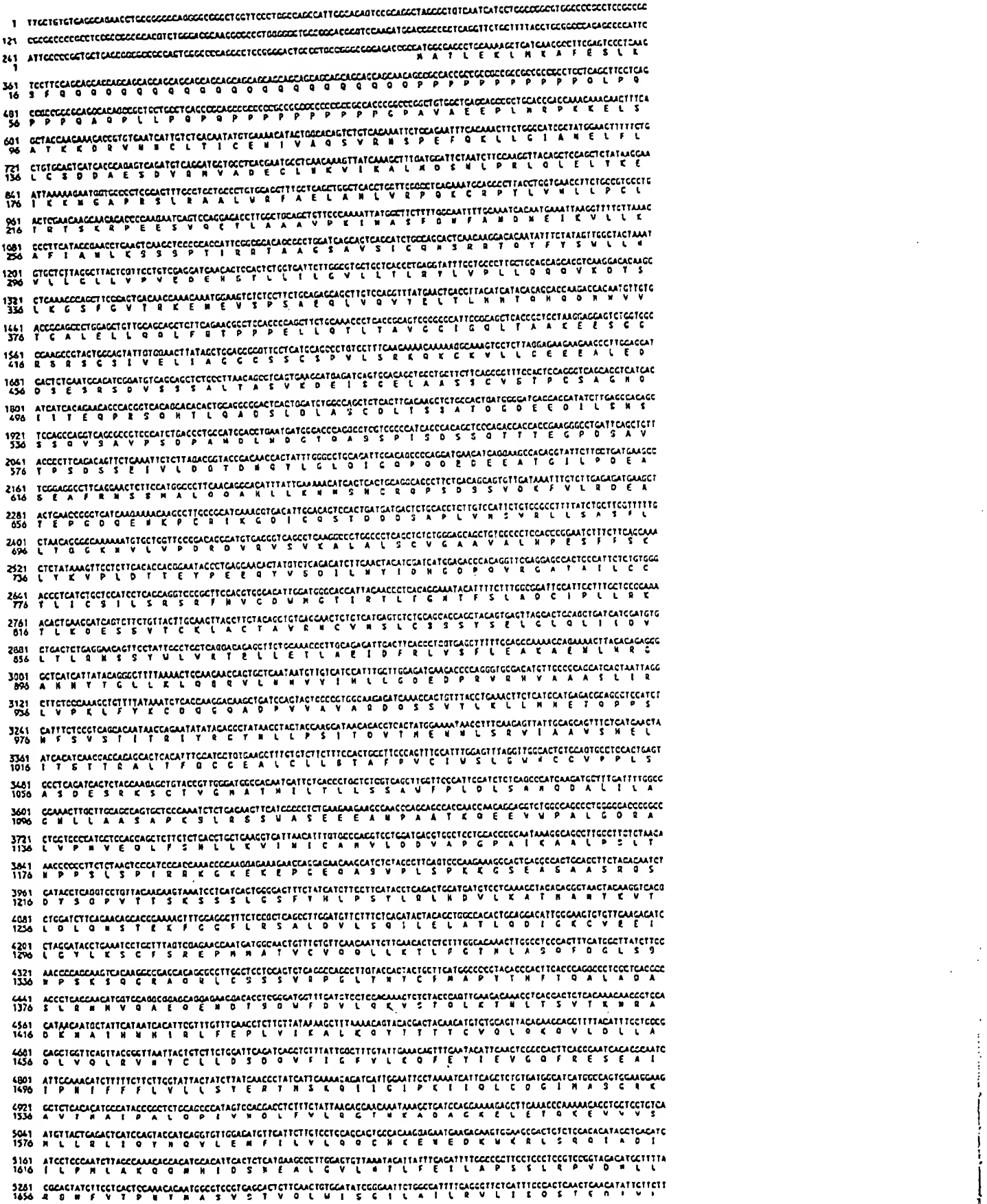


Figure 3. Schematic of cDNA Clones Defining the IT15 Transcript

Five cDNAs are represented under a schematic of the composite IT15 sequence. The thin line corresponds to untranslated regions. The thick line corresponds to coding sequence, assuming initiation of translation at the first Met codon in the open reading frame. Stars mark the positions of the following exon clones 5′ to 3′: DL83D3-8, DL83D3-1, DL228B6-3, DL228B6-5, DL228B6-13, DL69F7-3, DL178H4-6, DL118F5-U, and DL134B9-U4. The composite sequence was derived as follows. From 22 bases 3′ to the putative initiator Met ATG, the sequence was compiled from the cDNA clones and exons shown. There are 9 bases of sequence intervening between the 3′ end of IT16B and the 5′ end of IT15B. These were identified by PCR amplification of first-strand cDNA and sequencing of the PCR product. At the 5′ end of the composite sequence, the cDNA clone IT16C terminates 27 bases upstream of the (CAG)$_n$. However, when IT16C was identified, we had already generated genomic sequence surrounding the (CAG)$_n$ in an attempt to generate new polymorphisms. This sequence matched the IT16C sequence and extended it 337 bases upstream, including the apparent Met initiation codon.

Cell
974



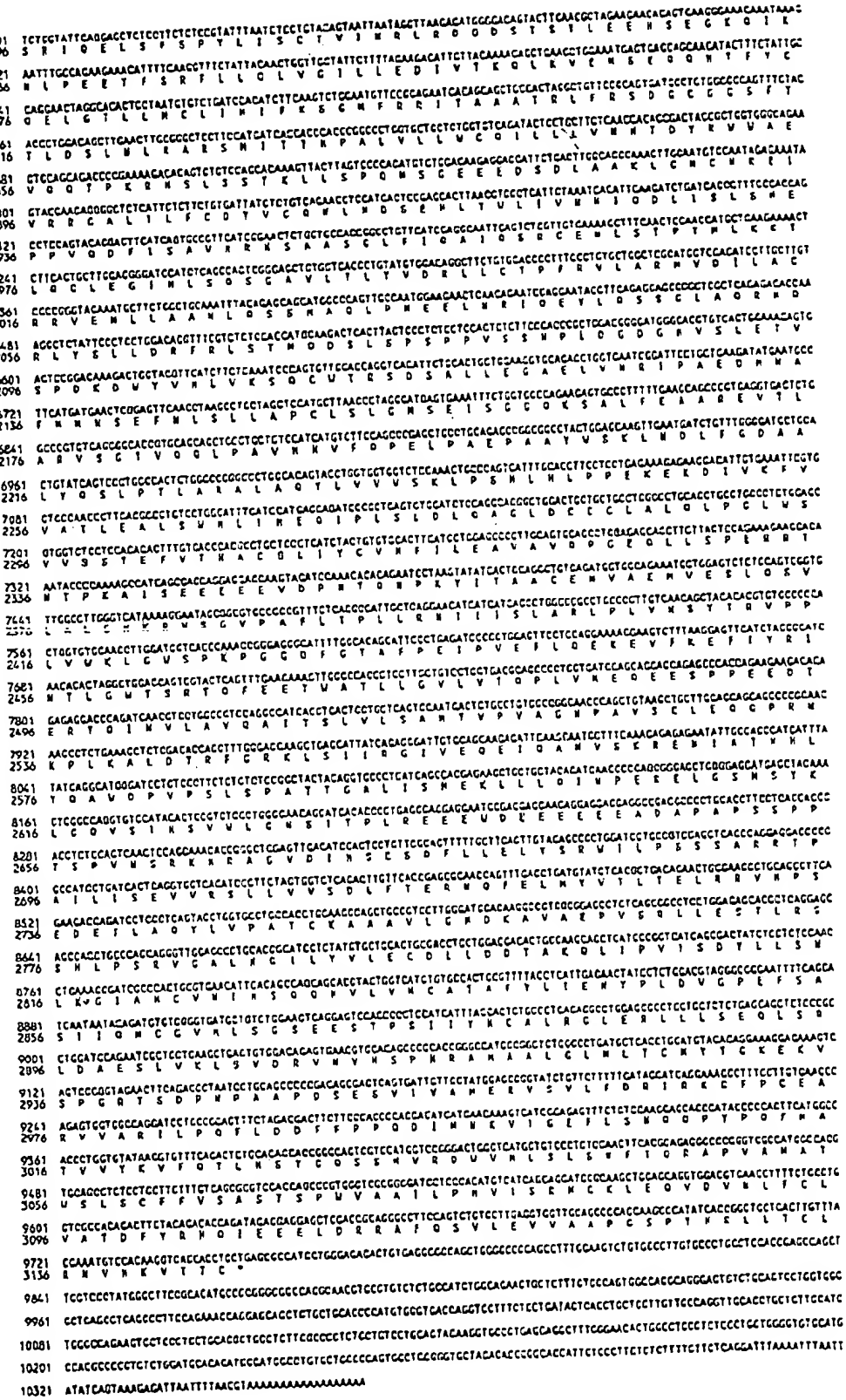


Figure 4. Composite Sequence of IT15

The composite DNA sequence of IT15 is shown. The predicted protein product is shown below the DNA sequence, based on the assumption that translation begins at the first in-frame methionine of the long open reading frame.

matic of the composite sequence derived as described in the figure legend. Figure 3 also displays the locations on the composite sequence of the nine trapped exon clones.

The composite sequence of IT15, containing the entire predicted coding sequence, spans 10,366 bases, including a tail of 18 A's as shown in Figure 4. An open reading frame of 9432 bases begins with a potential initiator methionine codon at base 316, located in the context of an optimal translation initiation sequence. An in-frame stop codon is located 240 bases upstream of this site. The protein product of IT15 is predicted to be a 348 kd protein containing 3144 amino acids. Although we have chosen the first Met codon in the long open reading frame as the probable initiator codon, we cannot exclude the possibility that translation does not actually begin at a more 3′ Met codon, producing a smaller protein.

Polymorphic Variation of the $(CAG)_n$ Trinucleotide Repeat

Near its 5′ end, the IT15 sequence contains 21 copies of the triplet CAG, encoding glutamine (Figure 5). When this sequence was compared with our collection of genomic sequences surrounding simple sequence repeats in 4p16.3, we found that normal cosmid L191F1 had 18 copies of the triplet, indicating that the $(CAG)_n$ repeat is polymorphic (Figure 5). We chose primers from the genomic sequence flanking the repeat to establish a polymerase chain reaction (PCR) assay for this variation. In the normal population, this simple sequence repeat polymorphism displays at least 17 discrete alleles, ranging from about 11 to 34 repeat units (Table 1). Ninety-eight percent of the 173 normal chromosomes tested contained repeat lengths between 11 and 24 repeats. Two chromosomes were detected in the 25–30 repeat range and 2 normal chromosomes had 33 and 34 repeats, respectively. The overall heterozygosity on normal chromosomes was 80%. We presume, based on sequence analysis of three clones, that the variation is based entirely on the $(CAG)_n$, but we cannot exclude the potential for variation of the smaller downstream $(CCG)_n$, which is also included in the PCR product.

Instability of the Trinucleotide Repeat on *HD* Chromosomes

Sequence analysis of cosmid GUS72-2130, derived from a chromosome with the major *HD* haplotype (see below), revealed 48 copies of the trinucleotide repeat, far more than the number of copies in the largest normal allele (Figure 5). When the PCR assay was applied to *HD* chromosomes, a pattern strikingly different from the normal variation was observed. *HD* heterozygotes contained one discrete allelic product in the normal size range and one PCR product of much larger size, suggesting that the $(CAG)_n$ repeat on *HD* chromosomes is expanded relative to normal chromosomes.

Figure 6 shows the patterns observed when we performed the PCR assay on lymphoblast DNA from a selected nuclear family in a large Venezuelan HD kindred. In this family, DNA marker analysis has shown previously that the *HD* chromosome was transmitted from the father

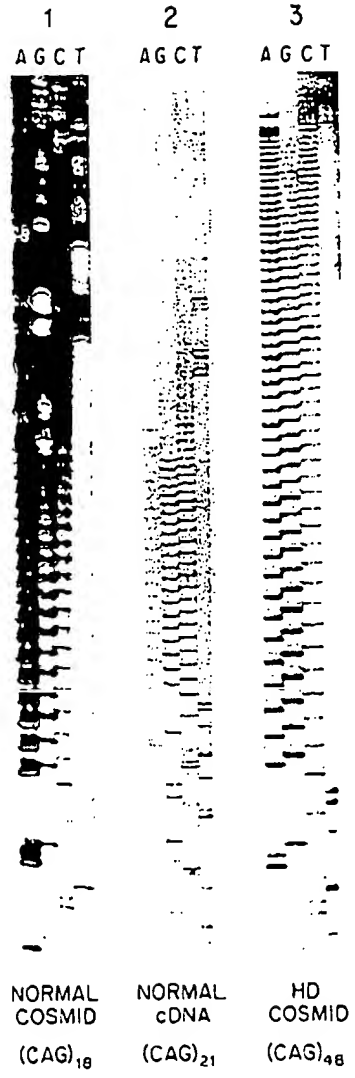


Figure 5. DNA Sequence Analysis of the $(CAG)_n$ Repeat

DNA sequence shown in panels 1, 2, and 3 demonstrates the variation in the $(CAG)_n$ repeat detected in normal cosmid L191F1 (1), cDNA IT16C (2), and *HD* cosmid GUS72-2130. Panels 1 and 3 were generated by direct sequencing of cosmid subclones using the primer 5′-GGCGGGAGACCGCCATGGCG-3′. Panel 2 was generated using the pBSKII T7 primer 5′-AATACGACTCACTATAG-3′.

Table 1. Comparison of HD and Normal Repeat Length

Range of Allele Sizes (Number of Repeats)	Normal Chromosomes		HD Chromosomes	
	Number	Frequency	Number	Frequency
≥48	0	0	44	0.59
42–47	0	0	30	0.41
30–41	2	0.01	0	0
25–30	2	0.01	0	0
≤24	169	0.98	0	0
Total	173	1.00	74	1.0

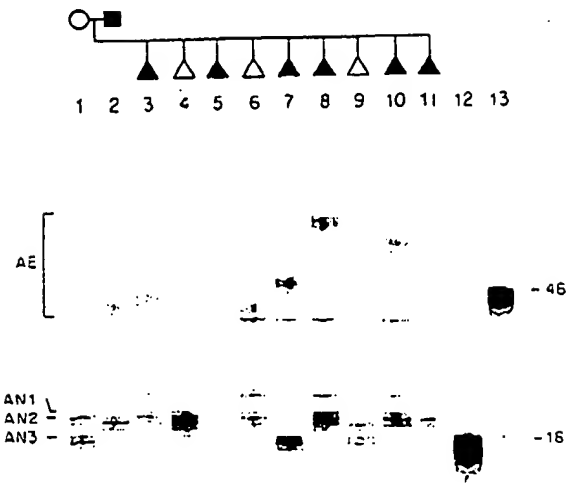


Figure 6. PCR Analysis of the (CAG)$_n$ Repeat in a Venezuelan HD Sibship with Some Offspring Displaying Juvenile Onset

Results of PCR analysis of a sibship in the Venezuelan HD pedigree are shown. Affected individuals are represented by closed symbols. Progeny are shown as triangles, and the birth order of some individuals has been changed for confidentiality. AN1, AN2, and AN3 mark the positions of the allelic products from normal chromosomes. AE marks the range of PCR products from the *HD* chromosome. The intensity of background constant bands, which represent a useful reference for comparison of the above PCR products, varies with slight differences in PCR conditions. The PCR products from cosmids L191F1 and GUS72-2130 are loaded in lanes 12 and 13 and have 18 and 48 CAG repeats, respectively.

(lane 2) to seven children (lanes 3, 5, 6, 7, 8, 10, and 11). The three normal chromosomes present in this mating yielded a PCR product in the normal size range (AN1, AN2, and AN3) that was inherited in a Mendelian fashion. The *HD* chromosome in the father yielded a diffuse, fuzzy PCR product slightly smaller than the 48 repeat product of our non-Venezuelan *HD* cosmid. Except for the DNA in lane 5, which did not PCR amplify, and in lane 11, which displayed only a single normal allele, each of the affected children's DNAs yielded a PCR product of a different size (AE), indicating instability of the *HD* chromosome (CAG)$_n$ repeat. Lane 6 contained an *HD*-specific product slightly smaller than or equal to that of the father's DNA. Lanes 3, 7, 10, and 8, respectively, contained *HD*-specific PCR products of progressively larger size. The absence of an *HD*-specific PCR product in lane 11 suggested that this child's DNA possessed a (CAG)$_n$ repeat that was too long to amplify efficiently. This was verified by Southern blot analysis in which the expanded *HD* allele was easily detected and estimated to contain up to 100 copies of the repeat. Notably, this child had juvenile onset of HD at the very early age of 2 years. The onset of HD in the father was when he was in his early 40s, typical of most adult HD patients in this population. The onset ages of the children represented by lanes 3, 7, 10, and 8 were 26, 25, 14, and 11 years, respectively, suggesting a rough correlation between age at onset of HD and the length of the (CAG)$_n$ repeat on the *HD* chromosome. In keeping with this trend, the offspring represented in lane 6 with the fewest repeats has reached adulthood without showing symptoms of the disorder.

Figure 7 shows PCR analysis for a second sibship from

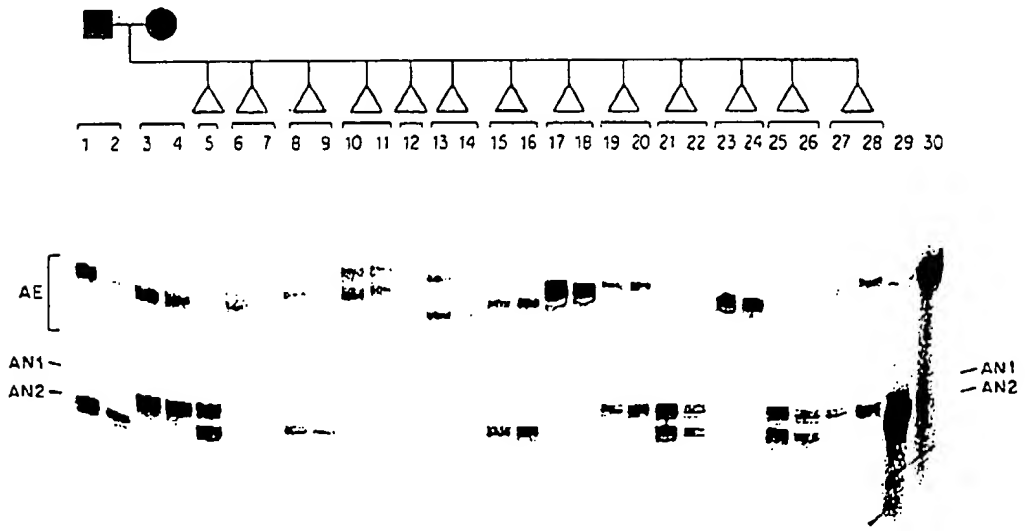


Figure 7. PCR Analysis of the (CAG)$_n$ Repeat in a Venezuelan HD Sibship with Offspring Homozygous for the Same *HD* Haplotype

Results of PCR analysis of a sibship from the Venezuelan HD pedigree in which both parents are affected by HD are shown. Progeny are shown as triangles and birth order has been altered for confidentiality. No HD diagnostic information is given to preserve the blind status of investigators in the Venezuelan Collaborative Group. AN1 and AN2 mark the positions of the allelic products from normal parental chromosomes. AE marks the range of PCR products from the *HD* chromosome. The PCR products from cosmids L191F1 and GUS72-2130 are loaded in lanes 29 and 30 and have 18 and 48 CAG repeats, respectively.

the Venezuelan pedigree, in which both parents are *HD* heterozygotes carrying the same *HD* chromosome based on DNA marker studies. Several of the offspring are *HD* homozygotes (lanes 6 and 7, 10 and 11, 13 and 14, 17 and 18, 23 and 24) as reported previously (Wexler et al., 1987). Each parent's DNA contained 1 allele in the normal range (AN1 and AN2), which was transmitted in a Mendelian fashion. The *HD*-specific products (AE) from the DNA of both parents and children were all much larger than the normal allelic products and also showed extensive variation in mean size. We have not provided a neurologic diagnosis for the offspring in this pedigree to maintain the blind status of investigators involved in the ongoing Venezuela HD Project, although age of onset again appears to parallel repeat length. Paired samples under many of the individual symbols represent independent lymphoblast lines initiated at least 1 year apart. The variance between paired samples was not as great as between the different individuals, suggesting that the major differences in size of the PCR products resulted from meiotic transmission. Of special note is the result obtained in lanes 13 and 14. This *HD* homozygote's DNA yielded one PCR product larger and one smaller than the *HD*-specific PCR products of both parents.

To date, we have tested 75 independent HD families, representing all different haplotypes reported by MacDonald et al. (1992) and a wide range of ethnic backgrounds. In all 75 cases, a PCR product larger than the normal size range was produced from the *HD* chromosome. The sizes of the *HD*-specific products ranged from 42 repeat copies to more than 66 copies, with a few individuals failing to yield a product because of the extreme length of the repeat. In these cases, Southern blot analysis revealed an increase in the length of an EcoRI fragment, with the largest allele approximating 100 copies of the repeat. Figure 8 shows the variation detected in members of an American family of Irish ancestry in which the major *HD* haplotype is segregating. Cosmid GUS72-2130 was cloned from the *HD* homozygous individual whose DNA was amplified in lane 2. As was observed in the Venezuelan HD pedigree (Figures 6 and 7), which segregates the disorder with a different 4p16.3 haplotype, the *HD*-specific PCR products for this family display considerable size variation.

New Mutations to *HD*?

The mutation rate in *HD* has been reported to be very low. To test whether the expansion of the $(CAG)_n$ repeat is the mechanism by which new *HD* mutations occur, we have examined two pedigrees with sporadic cases of HD in which intensive searching failed to reveal a family history of the disorder. In these cases, we gathered pedigree information sufficient to identify the same chromosomes in both the affected individual and unaffected relatives. Figure 9 shows the results of PCR analysis of the $(CAG)_n$ repeat in these families. The chromosomes in each family were assigned an arbitrary number based on typing for a large number of restriction fragment length polymorphism and simple sequence repeat markers in 4p16.3 defining distinct haplotypes; the presumed *HD* chromosome is starred.

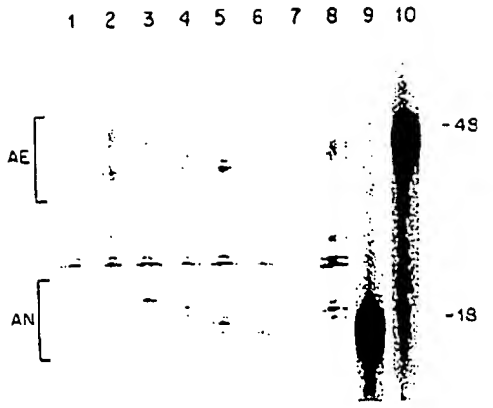


Figure 8. PCR Analysis of the $(CAG)_n$ Repeat in Members of an American Family with an Individual Homozygous for the Major *HD* Haplotype

Results of PCR analysis of members of an American family segregating the major *HD* haplotype. AN marks the range of normal alleles; AE marks the range of *HD* alleles. Lanes 1, 3, 4, 5, 7, and 8 represent PCR products from related *HD* heterozygotes. Lane 2 contains the PCR products from a member of the family homozygous for the same *HD* chromosome. Lane 6 contains PCR products from a normal individual. Pedigree relationships and affected status are not presented to preserve confidentiality. The PCR products from cosmids L191F1 and GUS72-2130 (which was derived from the individual represented in lane 2) are loaded in lanes 9 and 10 and have 18 and 48 CAG repeats, respectively.

In family 1, HD first appeared in individual II-3, who transmitted the disorder, along with chromosome 3*, to III-1. This same chromosome was present in II-2, an elderly unaffected individual. PCR analysis revealed that chromosome 3* from II-2 produced a PCR product at the extreme high end of the normal range (~36 CAG copies). However, the $(CAG)_n$ repeat on the same chromosome in II-3 and III-1 had undergone sequential expansions to ~44 and ~46 copies, respectively. A similar result was obtained in family 2, where the presumed new *HD* mutant III-2 had a considerably expanded repeat relative to the same chromosome in II-1 and III-1 (~49 versus ~33 CAG copies). In both families 1 and 2, the ultimate *HD* chromosome displays the marker haplotype characteristic of one-third of all *HD* chromosomes, suggesting that this haplotype may be predisposed to undergoing repeat expansion.

Discussion

The discovery of an expanded, unstable trinucleotide repeat on *HD* chromosomes suggests that the long-sought *HD* gene has at last been uncovered and that the disorder constitutes an example of a mutational mechanism that may prove quite common in human genetic disease. Elongation of a trinucleotide repeat sequence has been implicated previously as the cause of three quite different human disorders, the fragile X syndrome, myotonic dystrophy, and spino-bulbar muscular atrophy. Our initial observations of repeat expansion in HD indicate that this phenomenon shares features with each of these disorders.

In the fragile X syndrome, expression of a constellation of symptoms, including mental retardation and a fragile

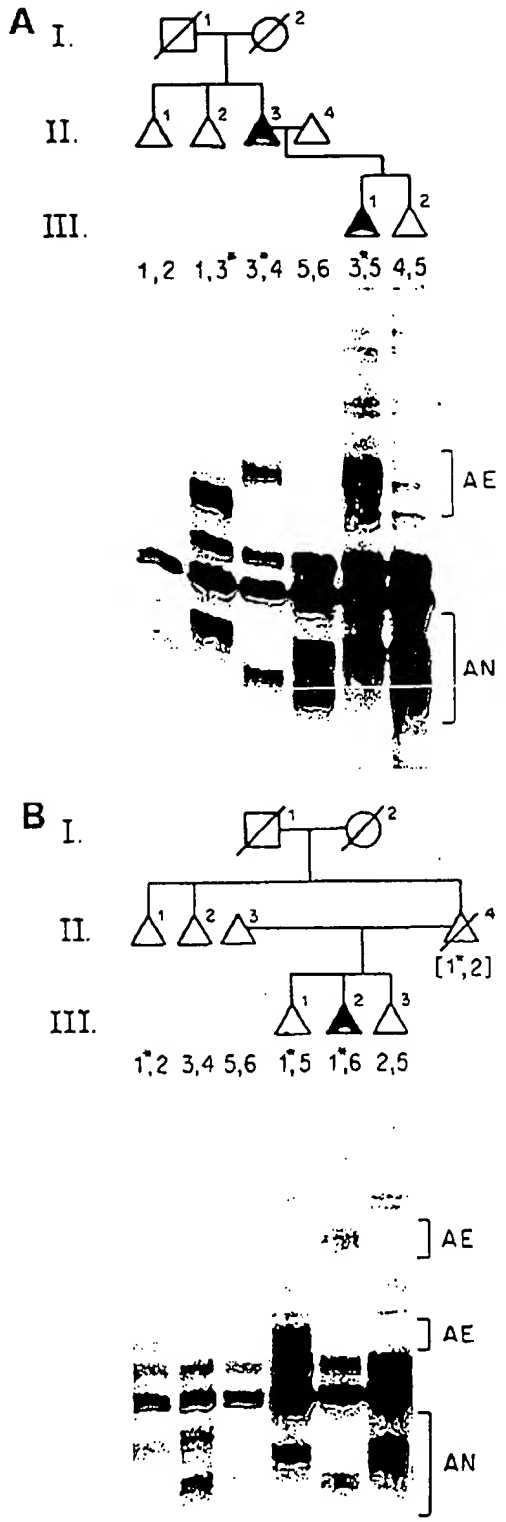


Figure 9. PCR Analysis of the $(CAG)_n$ Repeat in Two Families with a Supposed New Mutation Causing HD

Results of PCR analysis of two families in which sporadic HD cases representing putative new mutants are shown. Individuals in each pedigree are numbered by generation (roman numerals) and order in the pedigree. Triangles are used to protect confidentiality. Closed symbols indicate symptomatic individuals. The different chromosomes segregating in the pedigree have been distinguished by extensive typing with polymorphic markers in 4p16.3 and have been assigned arbitrary numbers shown above the gel lanes. The starred chromosomes (chromosome 3 in [A] and 1 in [B]) represent the presumed *HD* chromosome. AN denotes the range of normal alleles; AE denotes the range of alleles present in affected individuals and in their unaffected relatives bearing the same chromosome.

site at Xq27.3, is associated with expansion of a $(CGG)_n$ repeat thought to be in the 5′ untranslated region of the *FMR1* gene (Fu et al., 1991; Kremer et al., 1991; Verkerk et al., 1991). In myotonic dystrophy, a dominant disorder involving muscle weakness with myotonia that typically presents in early adulthood, the unstable trinucleotide repeat, $(CTG)_n$, is located in the 3′ untranslated region of the myotonin protein kinase gene (Aslanidis et al., 1992; Brook et al., 1992; Buxton et al., 1992; Fu et al., 1992; Harley et al., 1992a; Mahadevan et al., 1992). The unstable $(CAG)_n$ repeat in *HD* may be within the coding sequence of the IT15 gene, a feature shared with spino–bulbar muscular atrophy, an X-linked recessive adult-onset disorder of the motor neurons caused by expansion of a $(CAG)_n$ repeat in the coding sequence of the androgen receptor gene (LaSpada et al., 1991). The repeat length in both the fragile X syndrome and myotonic dystrophy tends to increase in successive generations, sometimes quite dramatically. Occasionally, decreases in the average repeat length are observed (Fu et al., 1991; Yu et al., 1992; Bruner et al., 1993). The *HD* trinucleotide repeat is also unstable, usually expanding when transmitted to the next generation, but contracting on occasion. In HD, as in the other disorders, change in copy number occurs in the absence of recombination. Compared with the fragile X syndrome, myotonic dystrophy, and HD, the instability of the disease allele in spino–bulbar muscular atrophy is more limited, and dramatic expansions of repeat length have not been seen (Biancalana et al., 1992).

Expansion of the repeat length in myotonic dystrophy is associated with a particular chromosomal haplotype, suggesting the existence of a primordial predisposing mutation (Harley et al., 1991, 1992a; Ashizawa, and Epstein, 1991). In the fragile X syndrome, there may be a limited number of ancestral mutations that predispose increases in trinucleotide repeat number (Richards et al., 1992; Oudet et al., 1993). The linkage disequilibrium analysis used to home in on IT15 indicates that there are several haplotypes associated with HD, but that at least one-third of *HD* chromosomes are ancestrally related (MacDonald et al., 1992). These data, combined with the reported low rate of new mutation to *HD* (Harper, 1992), suggest that expansion of the trinucleotide repeat may only occur on select chromosomes. Our analysis of two families in which new mutation was supposed to have occurred is consistent with the view that there may be particular normal chromosomes that have the capacity to undergo expansion of the repeat into the *HD* range. In each of these families, a chromosome with a $(CAG)_n$ repeat length in the upper end of the normal range was segregating on a chromosome

whose 4p16.3 haplotype matched the most common haplotype seen on *HD* chromosomes, and the clinical appearance of HD in these two cases was associated with expansion of the trinucleotide repeat.

The recent application of haplotype analysis to explore the linkage disequilibrium on *HD* chromosomes pointed to a portion of a 2.2 Mb candidate region defined by the majority of recombination events described in HD pedigrees (MacDonald et al., 1992). Previously, the search for the gene was confounded by three matings in which the genetic inheritance pattern was inconsistent with the remainder of the family (MacDonald et al., 1989b; Pritchard et al., 1992). These matings produced apparently affected HD individuals despite the inheritance of only normal alleles for markers throughout 4p16.3, effectively excluding inheritance of the *HD* chromosome present in the rest of the pedigree. Using our PCR assay, we have tested each of these families and find that, like other HD kindreds, an expanded allele generally segregated with *HD* in affected individuals of all three pedigrees. However, an expanded allele was not present in those specific individuals with the inconsistent 4p16.3 genotypes. Instead, these individuals displayed the normal alleles expected, based on analysis of other markers in 4p16.3. It is conceivable that these inconsistent individuals do not, in fact, have HD, but some other disorder. Alternatively, they might represent genetic mosaics in which the *HD* allele is more heavily represented and/or more expanded in brain tissue than in the lymphoblast DNA used for genotyping.

It can be expected that the capacity to monitor directly the size of the trinucleotide repeat in individuals "at risk" for HD will revolutionize preclinical testing for the disorder, eliminating the need for complicated linkage analyses, facilitating genetic counseling, and extending the applicability of presymptomatic and prenatal diagnosis to at risk individuals with no living affected relatives. We consider it of the utmost importance that the current internationally accepted guidelines and counseling protocols for testing those at risk continue to be observed, and that samples from unaffected relatives should not be tested inadvertently or without full consent. In our limited initial series of patients, there is an apparent correlation between repeat length and age of onset of the disease, reminiscent of that reported in myotonic dystrophy (Harley et al., 1992b; Tsilfidis et al., 1992). The largest *HD* trinucleotide repeat segments were found in juvenile onset cases, where there is a known preponderance of male transmission (Merrit et al., 1969). More detailed studies will be required to establish whether expansion of the repeat occurs preferentially in transmission from males. It will also be essential to perform a careful analysis of the extent, if any, of overlap between the range of repeat lengths in normal and HD individuals, to evaluate fully the relationship between age of onset and repeat length, and to examine the possibility of somatic variation in repeat length due to mitotic instability. These studies must be completed before the $(CAG)_n$ size is used to provide prognostic information to at risk HD individuals.

The expression of fragile X syndrome is associated with direct inactivation of the *FMR1* gene (Pieretti et al., 1991; DeBoulle et al., 1993). The recessive inheritance pattern of spino–bulbar muscular atrophy suggests that in this disorder an inactive gene product is produced. In myotonic dystrophy, the manner in which repeat expansion leads to the dominant disease phenotype is unknown. There are numerous possibilities for the mechanism of pathogenesis of the expanded trinucleotide repeat in *HD*. Since Wolf–Hirschhorn patients hemizygous for 4p16.3 do not display features of HD and IT15 mRNA is present in *HD* homozygotes, the expanded trinucleotide repeat does not cause simple inactivation of the gene containing it. The observation that the phenotype of *HD* is completely dominant, since homozygotes for the disease allele do not differ clinically from heterozygotes, has suggested that HD results from a gain-of-function mutation, in which either the mRNA product or the protein product of the disease allele would have some new property or would be expressed inappropriately (Wexler et al., 1987; Myers et al., 1989). If the expanded trinucleotide repeat were translated, the consequences on the protein product would be dramatic, increasing the length of the poly-glutamine stretch near the N-terminus. It is possible, however, that despite the presence of an upstream Met codon, the normal translational start occurs 3′ to the $(CAG)_n$ repeat and there is no poly-glutamine stretch in the protein product. In this case, the repeat would be in the 5′ untranslated region and might be expected to have its dominant effect at the mRNA level. The presence of an expanded repeat might directly alter regulation, localization, stability, or translatability of the mRNA containing it, and could indirectly affect its counterpart from the normal allele in *HD* heterozygotes. Other conceivable scenarios are that the presence of an expanded repeat might alter the effective translation start site for the *HD* transcript, thereby truncating the protein, or alter the transcription start site for the IT15 gene, disrupting control of mRNA expression. Finally, although the repeat is located within the IT15 transcript, the possibility that it leads to HD by virtue of an action on the expression of an adjacent gene cannot be excluded.

Despite this final caveat, we believe it most likely that the trinucleotide repeat expansion causes HD by its effect, either at the mRNA or protein level, on the expression and/or structure of the protein product of the IT15 gene, which we have named huntingtin. Outside of the region of the triplet repeat, the IT15 DNA sequence detected no significant similarity to any previously reported gene in the GenBank data base. Except for the stretches of glutamine and proline near the N-terminus, the amino acid sequence displayed no similarity to known proteins, providing no conspicuous clues to huntingtin's function. The poly-glutamine and poly-proline regions near the N-terminus indicate similarity to a large number of proteins that also contain long stretches of these amino acids. It is difficult to assess the significance of such similarities, although it is notable that many of these similarities are to DNA-binding proteins and that huntingtin does have a single leucine zipper motif at residue 1443. Huntingtin appears to be widely expressed, yet cell death in HD is confined to specific neurons in particular regions of the brain. Thus, with the mystery of the genetic basis of HD apparently solved,

defining the normal function of the huntingtin protein and delineating the mechanism whereby increased trinucleotide repeat length leads to the characteristic neuropathology of HD represent the next challenges in the effort to understand and to treat this devastating disorder.

Experimental Procedures

HD Cell Lines

Lymphoblast cell lines from HD families of varied ethnic backgrounds used for genetic linkage and disequilibrium studies (Conneally et al., 1989; MacDonald et al., 1992) have been established (Anderson and Gusella, 1984) in the Molecular Neurogenetics Unit, Massachusetts General Hospital, over the past 13 years. The Venezuelan HD pedigree is an extended kindred of over 12,000 members, in which all affected individuals have inherited the *HD* gene from a common founder (Gusella et al., 1983, 1984; Wexler et al., 1987).

DNA and RNA Blotting

DNA was prepared from cultured cells, and DNA blots were prepared and hybridized as described (Gusella et al., 1979, 1983). RNA was prepared and Northern blotting was performed as described by Taylor et al. (1992).

Construction of Cosmid Contig

The initial construction of the cosmid contig was by chromosome walking from cosmids L19 and BJ56 (Allitto et al., 1991; Lin et al., 1991). Two libraries were employed, a collection of Alu-positive cosmids from the reduced cell hybrid H39-8C10 (Whaley et al., 1991) and an arrayed flow-sorted chromosome 4 cosmid library (NM87545) provided by the Los Alamos National Laboratory. Walking was accomplished by hybridization of whole cosmid DNA, using suppression of repetitive and vector sequences, to robot-generated high density filter grids (Nizetic et al., 1991; Lehrach et al., 1990). Cosmids L1C2, L69F7, L22B6, and L83D3 were first identified by hybridization of yeast artificial chromosome clone YGA2 to the same arrayed library (Bates et al., 1992; Baxendale et al., 1991). *HD* cosmid GUS72-2130 was isolated by standard screening of a GUS72 cosmid library using a single-copy probe. Cosmid overlaps were confirmed by a combination of clone to clone and clone to genomic hybridizations, single-copy probe hybridizations, and restriction mapping.

cDNA Isolation and Characterization

Exon probes were isolated and cloned as described (Buckler et al., 1991). Exon probes and cDNAs were used to screen human λZAPII cDNA libraries constructed from adult frontal cortex, fetal brain, adenovirus-transformed retinal cell line RCA, and liver RNA. cDNA clones, PCR products, and trapped exons were sequenced as described (Sanger et al., 1977). Direct cosmid sequencing was performed as described (McClatchey et al., 1992). Data base searches were performed using the BLAST network service of the National Center for Biotechnology Information (Altschul et al., 1990).

PCR Assay of the $(CAG)_n$ Repeat

Genomic primers flanking the $(CAG)_n$ repeat are 5′-ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC-3′ and 5′-AAACTCACGGTCGGTGCAGCGGCTCCTCAG-3′. PCR amplification was performed in a reaction volume of 25 μl using 50 ng of genomic DNA, 5 μg of each primer, 10 mM Tris (pH 8.3), 5 mM KCl, 2 mM $MgCl_2$, 200 μM (each) dNTPs, 10% dimethylsulfoxide, 0.1 U of Perfectmatch (Stratagene), 2.5 μCi of [^{32}P]dCTP (Amersham), and 1.25 U of Taq polymerase (Boehringer Mannheim). After heating to 94°C for 1.5 min, the reaction mix was cycled according to the following program: 40 cycles of 1 min at 94°C, 1 min at 60°C, 2 min at 72°C. Five microliters of each PCR was diluted with an equal volume of 95% formamide loading dye and heat denatured for 2 min at 95°C. The products were resolved on 5% denaturing polyacrylamide gels. The PCR product from this reaction using cosmid L191F1 (CAG_{18}) as the template was 247 bp. Allele sizes were estimated relative to a DNA sequencing ladder, the PCR products from sequenced cosmids, and the invariant background bands often present on the gel. Estimates of allelic variation were obtained by typing unrelated individuals of largely Western European ancestry, who were normal parents of affected HD individuals from various pedigrees.

Acknowledgments

We thank the many investigators who have supplied blood samples from HD pedigrees, particularly the Venezuela Huntington's Disease Collaborative Group, and the many investigators who have over the years participated in or contributed to this project. We are also extremely grateful to the HD families themselves for supporting and participating in this research effort. We thank P. M. Conneally, G. Evans, M. Frangione, C. Gilliam, R. Horvitz, L. Moyzis, R. Mulligan, A. Novelletto, A. Tobin, and L. Zipursky for helpful discussions. This work was supported by National Institutes of Health grants NS16367 (Huntington's Disease Center Without Walls), NS22031, and NS25631 and by grants from Bristol-Myers Squibb, Inc., the Hereditary Disease Foundation Collaborative Research Agreement, the Joan and William A. Schroyer/Merrill Lynch Fund to Cure Huntington's Disease of the Hereditary Disease Foundation, the Huntington's Disease Society of America, the Foundation for the Care and Cure of Huntington's Disease, the W. M. Keck Foundation, the William J. Matheson Foundation, the Bay Foundation, The Charles A. Dana Foundation, the Medical Research Council (England), the Welsh Office, and the Wellcome Trust. Fellowship support was provided to C. M. A. by the Andrew B. Cogan Fellowship of the Hereditary Disease Foundation. Other postdoctoral fellowships to various investigators were generously provided by the Hereditary Disease Foundation, the Huntington's Disease Society of America, and the Huntington's Society of Canada.

Received February 26, 1993; revised March 4, 1993.

References

Allitto, B. A., MacDonald, M. E., Bucan, M., Richards, J., Romano, D., Whaley, W. L., Falcone, B., Ianazzi, J., Wexler, N. S., Wasmuth, J. J., Collins, F. S., Lehrach, H., Haines, J. L., and Gusella, J. F. (1991). Increased recombination adjacent to the Huntington's disease-linked *D4S10* marker. Genomics *9*, 104–112.

Altherr, M. R., Plummer, S., Bates, G., MacDonald, M., Taylor, S., Lehrach, T. H., Frischauf, A. M., Gusella, J. F., Boehnke, M., and Wasmuth, J. J. (1992). Radiation hybrid map spanning the Huntington disease gene region of chromosome 4. Genomics *13*, 1040–1046.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Ambrose, C., James, M., Barnes, G., Lin, C., Bates, G., Altherr, M., Duyao, M., Groot, N., Church, D., Wasmuth, J. J., Lehrach, H., Housman, D., Buckler, A., Gusella, J. F., and MacDonald, M. E. (1992). A novel G protein-coupled receptor kinase cloned from 4p16.3. Hum. Mol. Genet. *1*, 697–703.

Anderson, M. A., and Gusella, J. F. (1984). Use of cyclosporin A in establishing Epstein–Barr virus-transformed human lymphoblastoid cell lines. In Vitro *20*, 856–858.

Ashizawa, T., and Epstein, H. F. (1991). Ethnic distribution of myotonic dystrophy gene. Lancet *338*, 642–643.

Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N. G. M., Vooijs, M., Buxton, J., Johnson, K., Smeets, H. J. M., Lennon, G. G., Carrano, A. V., Korneluk, R. G., Wieringa, B., and de Jong, P. J. (1992). Cloning of the essential myotonic dystrophy region and mapping of the putative defect. Nature *355*, 548–551.

Bates, G. P., MacDonald, M. E., Baxendale, S., Sedlacek, Z., Youngman, S., Romano, D., Whaley, W. L., Allitto, B. A., Poustka, A., Gusella, J. F., and Lehrach, H. (1990). A YAC telomere clone spanning a possible location of the Huntington's disease gene. Am. J. Hum. Genet. *46*, 762–775.

Bates, G. P., MacDonald, M. E., Baxendale, S., Youngman, S., Lin, C., Whaley, W. L., Wasmuth, J. J., Gusella, J. F., and Lehrach, H. (1991). Defined physical limits of the Huntington disease gene candidate region. Am. J. Hum. Genet. *49*, 7–16.

Bates, G. P., Valdes, J., Hummerich, H., Baxendale, S., Le Paslier, D. L., Monaco, A. P., Tagle, D., MacDonald, M. E., Altherr, M., Ross,

M., Brownstein, B. H., Bentley, D., Wasmuth, J. J., Gusella, J. F., Cohen, D., Collins, F., and Lehrach, H. (1992). Characterization of a yeast artificial chromosome contig spanning the Huntington disease gene candidate region. Nature Genet. *1*, 180–187.

Baxendale, S., Bates, G. P., MacDonald, M. E., Gusella, J. F., and Lehrach, H. (1991). The direct screening of cosmid libraries with yeast artificial chromosomes. Nucl. Acids Res. *19*, 6851.

Biancalana, V., Serville, F., Pommier, J., Julien, J., Hanauer, A., and Mandel, J. L. (1992). Moderate instability of the trinucleotide repeat in spino–bulbar muscular atrophy. Hum. Mol. Genet. *1*, 255–258.

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J.-P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J., and Housman, D. E. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell *68*, 799–808.

Bruner, H. G., Jansen, G., Nillesen, W., Nelen, M. R., DeDie, C. E. M., Howeler, C. J., van Oost, B. A., Wieringa, B., Ropers, H. H., and Smeets, H. J. M. (1993). Reverse mutation in myotonic dystrophy. N. Engl. J. Med. *328*, 476–480.

Bucan, M., Zimmer, M., Whaley, W. L., Poustka, A., Youngman, S., Allitto, B. A., Ormondroyd, E., Smith, B., Pohl, T. M., MacDonald, M., Bates, G., Richards, J., Volinia, S., Gilliam, T. C., Sedlacek, Z., Collins, F. S., Wasmuth, J. J., Shaw, D. J., Gusella, J. F., Frischauf, A. M., and Lehrach, H. (1990). Physical maps of 4p16.3, the area expected to contain the Huntington's disease mutation. Genomics *6*, 1–15.

Buckler, A. J., Chang, D. D., Graw, S. L., Brook, J. D., Haber, D. A., Sharp, P. A., and Housman, D. E. (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc. Natl. Acad. Sci. USA *88*, 4005–4009.

Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B., Williamson, R., and Johnson, K. (1992). Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. Nature *355*, 547–548.

Conneally, P. M., Haines, J. L., Tanzi, R. E., Wexler, N. S., Penchaszadeh, G. K., Harper, P. S., Folstein, S. E., Cassiman, J. J., Myers, R. H., Young, A. B., Hayden, M. R., Falek, A., Tolosa, E. S., Crespi, S., Di Maio, L., Holmgren, G., Anvret, M., Kanazawa, I., and Gusella, J. F. (1989). Huntington disease: no evidence for locus heterogeneity. Genomics *5*, 304–308.

DeBoulle, K., Verkerk, A. J. M. H., Reyniers, E., Vits, L., Hendrickx, J., VanRoy, B., VanDenBos, F., deGraaff, E., Oostra, B. A., and Willems, P. J. (1993). A point mutation in the *FMR-1* gene associated with fragile X mental retardation. Nature Genet. *3*, 31–35.

Doucette-Stamm, L. A., Riba, L., Handelin, B., Olfilippantonio, M., Ward, D. C., Wasmuth, J. J., Gusella, J. F., and Housman, D. E. (1991). Generation and characterization of Goss-Harris hybrids of human chromosome 4. Somat. Cell Mol. Genet. *17*, 471–480.

Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkerk, A. J. M. H., Holden, J. J. A., Fenwick, R. G., Jr., Warren, S. T., Oostra, B. A., Nelson, D. L., and Caskey, C. T. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell *67*, 1047–1058.

Fu, Y.-H., Pizzuti, A., Fenwick, R. G., King, J., Jr., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., DeJong, P., Wieringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F., and Caskey, C. T. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science *255*, 1256–1259.

Gusella, J. F. (1989). Location cloning strategy for characterizing genetic defects in Huntington's disease and Alzheimer's disease. FASEB J. *3*, 2036–2041.

Gusella, J. F. (1991). Huntington's disease. Adv. Hum. Genet. *20*, 125–151.

Gusella, J. F., Varsanyi-Breiner, A., Kao, F. T., Jones, C., Puck, T. T., Keys, C., Orkin, S., and Housman, D. E. (1979). Precise localization of the human β-globin gene complex on chromosome 11. Proc. Natl. Acad. Sci. USA *76*, 5239–5243.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E., and Martin, J. B. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234–238.

Gusella, J. F., Tanzi, R. E., Anderson, M. A., Hobbs, W., Gibbons, K., Raschtchian, R., Gilliam, T. C., Wallace, M. R., Wexler, N. S., and Conneally, P. M. (1984). DNA markers for nervous system diseases. Science *225*, 1320–1326.

Harley, H. G., Brook, J. D., Floyd, J., Rundle, S. A., Crow, S., Walsh, K. V., Thibault, M. C., Harper, P. S., and Shaw, D. J. (1991). Detection of linkage disequilibrium between the myotonic dystrophy locus and a new polymorphic DNA marker. Am. J. Hum. Genet. *49*, 68–75.

Harley, H. G., Brook, J. D., Rundle, S. A., Crow, S., Reardon, W., Buckler, A. J., Harper, P. S., and Housman, D. E. (1992a). Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. Nature *355*, 545–546.

Harley, H. G., Rundle, S. A., Reardon, W., Myring, J., Crow, S., Brook, J. D., Harper, P. S., and Shaw, D. J. (1992b). Unstable DNA sequence in myotonic dystrophy. Lancet *339*, 1125–1128.

Harper, P. S. (1992). The epidemiology of Huntington's disease. J. Med. Genet. *89*, 365–376.

Harper, P. S., Morris, M. J., Quarrell, O., Shaw, D. J., Tyler, A., and Youngman, S. (1991). Huntington's Disease (Philadelphia: W. B. Saunders).

Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R., and Richards, R. I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science *252*, 1711–1714.

LaSpada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fishbeck, H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature *352*, 77–79.

Lehrach, H., Drmanac, R., Hoheisel, J., Larin, Z., Lennon, G., Nizetic, D., Monaco, A., Zehetner, G., and Poustka, A. (1990). Hybridisation fingerprinting in genome mapping and sequencing. In Genome Analysis: Genetic and Physical Mapping, Volume 1, K. E. Davies and S. M. Tighlman, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press), pp. 39–81.

Lin, C. S., Altherr, M., Bates, G., Whaley, W. L., Read, A. P., Harris, R., Lehrach, H., Wasmuth, J. J., Gusella, J. F., and MacDonald, M. E. (1991). New DNA markers in the Huntington's disease gene candidate region. Somat. Cell Mol. Genet. *17*, 481–488.

MacDonald, M. E., Cheng, S. V., Zimmer, M., Haines, J. L., Poustka, A. M., Allitto, B. A., Smith, B., Whaley, W. L., Romano, D., Jagadeesh, J., Lehrach, H., Wasmuth, J. J., Frischauf, A. M., and Gusella, J. F. (1989a). Clustering of multi-allele DNA markers near the Huntington's disease gene. J. Clin. Invest. *84*, 1013–1016.

MacDonald, M. E., Haines, J. L., Zimmer, M., Cheng, S. V., Youngman, S., Whaley, W. L., Bucan, W. L., Allitto, B. A., Smith, B., Leavitt, J., Poustka, A. M., Harper, P., Lehrach, H., Wasmuth, J. J., Frischauf, A. M., and Gusella, J. F. (1989b). Recombination events suggest possible locations for the Huntington's disease gene. Neuron *3*, 183–190.

MacDonald, M. E., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W. L., Lehrach, H., Wasmuth, J., and Gusella, J. F. (1991). Complex patterns of linkage disequilibrium in the Huntington disease region. Am. J. Hum. Genet. *49*, 723–734.

MacDonald, M. E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., Lehrach, H., Collins, F. S., Wasmuth, J. J., Frontali, M., and Gusella, J. F. (1992). The Huntington's disease candidate region exhibits many different haplotypes. Nature Genet. *1*, 99–103.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-MacDonald, J., DeJong, P. J., Wieringa, B., and Korneluk, G. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. Science *255*, 1253–1255.

Martin, J. B., and Gusella, J. F. (1986). Huntington's disease: pathogenesis and management. N. Engl. J. Med. *315*, 1267–1276.

McClatchey, A. I., Lin, C. S., Wang, J., Hoffman, E. P., Rojas, C., and

Gusella, J. F. (1992). The genomic structure of the human skeletal muscle sodium channel gene. Hum. Mol. Genet. *1*, 521–527.

Merrit, A. D., Conneally, P. M., Rahman, N. F., and Drew, A. L. (1969). Juvenile Huntington's chorea. In Progress in Neurogenetics, A. Barbeau and J. R. Brunette, eds. (Amsterdam: Excerpta Medica), pp. 645–650.

Myers, R. H., Leavitt, J., Farrer, L. A., Jagadeesh, J., McFarlane, H., Mark, R. J., and Gusella, J. F. (1989). Homozygote for Huntington's disease. Am. J. Hum. Genet. *45*, 615–618.

Nizetic, D., Zehetner, G., Monaco, A., Gellen, L., Young, B. D., and Lehrach, H. (1991). Construction, arraying, and high density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. Proc. Natl. Acad. Sci. USA *88*, 3233–3237.

Oudet, C., Mornet, E., Serre, J. L., Thomas, F., Lentes-Zengerling, S., Kretz, C., Deluchat, C., Tejada, I., Boue, J., Boue, A., and Mandel, J. L. (1993). Linkage disequilibrium between the fragile X mutation and two closely linked CA repeats suggests that fragile X chromosomes are derived from a small number of founder chromosomes. Am. J. Hum. Genet. *52*, 297–304.

Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T., and Nelson, D. L. (1991). Absence of expression of the *FMR-1* gene in fragile X syndrome. Cell *66*, 817–822.

Pohl, T. M., Zimmer, M., MacDonald, M. E., Smith, B., Bucan, M., Poustka, A., Vollnia, S., Searle, S., Zehetner, G., Wasmuth, J. J., Gusella, J., Lehrach, H., and Frischauf, A. M. (1988). Construction of a NotI linking library and isolation of new markers close to the Huntington's disease gene. Nucl. Acids Res. *16*, 9185–9198.

Pritchard, C., Zhu, N., Zuo, J., Bull, L., Pericak-Vance, M. A., Roses, A. D., Milatovich, A., Francke, U., Cox, D. R., and Myers, R. M. (1992). Recombination of 4p16 DNA markers in an unusual family with Huntington disease. Am. J. Hum. Genet. *50*, 1218–1230.

Richards, R. I., Holman, K., Friend, K., Kremer, E., Hillen, D., Staples, A., Brown, W. T., Goonewardena, P., Tarleton, J., Schwartz, C., and Sutherland, G. R. (1992). Fragile X syndrome: evidence of founder chromosomes. Nature Genet. *1*, 257–260.

Sanger, T., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-termination inhibitors. Proc. Natl. Acad. Sci. USA *74*, 5463–5467.

Snell, R. G., Lazarou, L., Youngman, S., Quarrell, O. W. J., Wasmuth, J. J., Shaw, D. J., and Harper, P. S. (1989). Linkage disequilibrium in Huntington's disease: an improved localization for the gene. J. Med. Genet. *26*, 673–675.

Snell, R. G., Thompson, L. M., Tagle, D. A., Holloway, T. L., Barnes, G., Harley, H. G., Sandkuijl, L. A., MacDonald, M. E., Collins, F. S., Gusella, J. F., Harper, P. S., and Shaw, D. J. (1992). A recombination event that redefines the Huntington disease region. Am. J. Hum. Genet. *51*, 357–362.

Suthers, G. K., Huson, S. M., and Davies, K. E. (1992). Instability versus predictability: the molecular diagnosis of myotonic dystrophy. J. Med. Genet. *29*, 761–765.

Taylor, S. A. M., Snell, R. G., Buckler, A., Ambrose, C., Duyao, M., Church, D., Lin, C. S., Altherr, M., Bates, G. P., Groot, N., Barnes, G., Shaw, D. J., Lehrach, H., Wasmuth, J. J., Harper, P. S., Housman, D. E., MacDonald, M. E., and Gusella, J. F. (1992). Cloning of the α-adducin gene from the Huntington's disease candidate region of chromosome 4 by exon amplification. Nature Genet. *2*, 223–227.

Theilman, J., Kanani, S., Shiang, R., Robbins, C., Quarrell, O., Huggins, M., Hedrick, A., and Hayden, M. R. (1989). Non-random association between alleles detected at *D4S95* and *D4S98* and the Huntington's disease gene. J. Med. Genet. *26*, 676–681.

Thompson, L. M., Plummer, S., Schalling, M., Altherr, M. R., Gusella, J. F., Housman, D. E., and Wasmuth, J. J. (1991). A gene encoding a fibroblast growth factor receptor isolated from the Huntington disease gene region of human chromosome 4. Genomics *11*, 1133–1142.

Tsilfidis, C., McKenzie, A. E., Mettler, G., Barcelo, J., and Korneluk, R. G. (1992). Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy. Nature Genet. *1*, 192–195.

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, R., Eussen, B. E., van Ommen, G.-J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A., and Warren, S. T. (1991). Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell *65*, 905–914.

Wexler, N. S., Young, A. B., Tanzi, R. E., Travers, H., Starosta-Rubenstein, S., Penney, J. B., Snodgrass, S. R., Shoulson, I., Gomez, F., Ramos-Arroyo, M. A., Penchaszadeh, G., Moreno, R., Gibbons, K., Faryniarz, A., Hobbs, W., Anderson, M. A., Bonilla, E., Conneally, P. M., and Gusella, J. F. (1987). Homozygotes for Huntington's disease. Nature *326*, 194–197.

Whaley, W. L., Bates, G. P., Novelleto, A., Sedlacek, Z., Cheng, S., Romano, D., Ormondroyd, E., Allitto, B. A., Lin, C., Youngman, S., Baxendale, S., Bucan, M., Altherr, M., Wasmuth, J., Wexler, N. S., Frontali, M., Frischauf, A. M., Lehrach, H., MacDonald, M. E., and Gusella, J. F. (1991). Mapping of cosmid clones in the Huntington's disease region of chromosome 4. Somat. Cell Mol. Genet. *17*, 83–91.

Youngman, S., Bates, G. P., Williams, S., McClatchey, A. I., Baxendale, S., Sedlacek, Z., Altherr, M., Wasmuth, J. J., MacDonald, M. E., Gusella, J. F., and Lehrach, H. (1992). The telomeric 60 kb of chromosome arm 4p is homologous to telomeric regions on 13p, 15p, 21p, and 22p. Genomics *14*, 350–356.

Yu, S., Mulley, J., Loesch, D., Turner, G., Donnelly, A., Gedeon, A., Hillen, D., Kremer, E., Lynch, M., Pritchard, M., Sutherland, G. R., and Richards, R. I. (1992). Fragile-X syndrome: unique genetics of the heritable unstable element. Am. J. Hum. Genet. *50*, 968–980.

GenBank Accession Number

The accession number for the sequence reported in this paper is L12392.